AD_____

Award Number: DAMD17-97-1-7058

TITLE: Development of an Integrated Program of Health Related Quality of Life Research for the National Surgical Adjuvant Breast and Bowel Project

PRINCIPAL INVESTIGATOR: Richard D. Day, Ph.D.

CONTRACTING ORGANIZATION: University of Pittsburgh
Pittsburgh, Pennsylvania 15260

REPORT DATE: September 2001

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 2001	Annual (1 Sep 00 - 31 Aug 01)

4. TITLE AND SUBTITLE
Development of an Integrated Program of Health Related Quality of Life Research for the National Surgical Adjuvant Breast and Bowel Project

5. FUNDING NUMBERS
DAMD17-97-1-7058

6. AUTHOR(S)
Richard D. Day, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pittsburgh
Pittsburgh, Pennsylvania 15260

E-Mail: rdfac@vms.cis.pitt.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

This Career Development Award (CDA) was specifically intended to support Dr. Day in the development of a Health-Related Quality of Life Program (HRQL) for the National Surgical Adjuvant Breast and Bowel Project (NSABP). There are now a total of 3 ongoing NSABP HRQL studies in the process of data collection (P-2, B-32, B-33). Two studies which were underway at the last reporting date have now been closed to recruitment (B30, C-07). For the past 12 months of the CDA, Dr. Day has had reduced responsibility for day-to-day monitoring of ongoing protocols and has served as a consultant for study development. This has permitted him to devote an increased amount of time in months 37-48 of the grant to the completion of manuscripts and reports. A no cost 12 month extension of the award has been approved enabling Dr. Day to prepare additional analyses and publications on NSABP quality-of-life data.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Breast Cancer	103
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Career Development Award:

Development of an Integrated Program of Health-Related Quality of Life Research for the National Surgical Adjuvant Breast and Bowel Project

Richard Day, Ph.D.
Department of Biostatistics
University of Pittsburgh

Fourth Annual Progress Report
September 1, 2000 to August 31, 2001

1. Introduction

This Career Development Award (CDA) was specifically intended to support Dr. Day in the development of a Health-Related Quality of Life Program (HRQL) for the National Surgical Adjuvant Breast and Bowel Project (NSABP). Specific aims proposed for the CDA included: (a) Design and implementation of new HRQL components for planned NSABP treatment and prevention trials; (b) testing and implementation of data collection methods to be used in treatment and prevention trials; (c) analysis of HRQL data collected in the NSABP prevention and treatment trials; (d) refinement and extension of HRQL methods to analyze the data from new treatment and prevention studies; (e) enhancement of minority participation in NSABP trials. Work completed during the third 12 months of Dr. Day's CDA will be summarized in terms of these aims.

2. Body

2.1 Design and implementation of new HRQL breast cancer components for planned NSABP treatment and prevention trials.

Update of Health-Related Quality of Life (HRQL) protocols implemented in as part of Dr. Day's CDA:

a. **Protocol no. B-33 – *A Randomized, Placebo Controlled, Double-Blinded Trial Evaluating the Effect of Exemestane in Stage I and II Post-Menopausal Breast Cancer Patients Completing at least Five Years of Tamoxifen Therapy.*** The HRQL component of this protocol will involve 300 patients randomized to each arm of the trial. It is expected that examestane will have significant effects on the patients' quality of life. As an extremely effective aromatase inhibitor, the resulting lack of estrogen is expected to be associated with an increase in the frequency and intensity of menopausal symptoms. It is expected that any increase in symptoms will occur relatively quickly after the initiation of examestane and will remain stable as long as the medication is taken. The instrument selected for use in this study is the Menopause Specific Quality of Life Questionnaire. Use of this instrument will permit a comparison of the B-33 findings to a study of letrozole being carried out by the NCI of Canada. **Current Status (09/01/01): This trial has just opened this month and 3 patients have so far been recruited into the HRQL study.**

b. **Protocol no. B-32 – *A Randomized, Phase III Clinical Trial to Compare Sentinel Node Resection to Conventional Axillary Dissection in Clinically Node Negative Breast Cancer Patients*.** Axillary lymph node resection is generally performed on women with operable breast cancer in order to aid in the determination of staging, need for adjuvant therapy, and regional control. Although rarely life-threatening, axillary lymph node dissection is associated with significant morbidity. Patients often experience reduced mobility of the shoulder and require physical therapy to regain full function of the upper extremity. Lymphedema has been reported by 30% of women who have had conservative breast surgery with axillary dissection. In B-32, axillary dissection is compared to a new surgical method, sentinel node resection. Sentinel node resection usually requires the identification and removal of a single lymph node (or a small number of nodes) from the axilla. Because the procedure involves much less extensive surgery than traditional axillary dissection, we expect that it will result in less morbidity and allow for more rapid return to normal activity, with fewer long-term sequale.

This study involved the development and testing of a new, self-administered HRQL questionnaire (Physical Functioning Questionnaire). In addition, we are using a general QOL Rating Scale (0-10), anchored by death and perfect health. This scale has been used in previous NSABP studies. Approximately 325 patients from each study arm, stratified by type of operation (lumpectomy/ mastectomy) will be included in this study. . **Current Status (09/01/01): The HRQL component of this trial is now open and 19 patients have so far completed the baseline examination.**

c. **Protocol no. B-30 – *A Three Arm Randomized Trial to Compare Adjuvant Adriamycin and Cyclophosphamide Followed by Taxotere (AC-T); Adriamycin and Taxotere (AT); and Adriamycin, Taxotere and Cyclophosphamide (ATC) in Breast Cancer Patients with Positive Axillary Lymph Notes.*** Quality of life measures included in B-30 questionnaire are the Functional Assessment of Cancer Therapy-Breast (FACT-B), a treatment specific Symptom Checklist (SCL), the SF-36 Vitality Scale, and an overall HRQL rating scale. Two additional questionnaires (Baseline and Follow-Up Menstrual History Questionnaires) were developed and tested in order to measure ovarian damage occurring as a long-term sequela of adjuvant chemotherapy. . **Current Status (09/01/01): This trial was closed to recruitment on 7/20/2001 and includes 2107 patients in the HRQL study.**

d. **Protocol no. P-2 – *Study of Tamoxifen and Raloxifene (STAR).*** This is the new NSABP prevention study following on the positive results of the P-1 (Breast Cancer Prevention Trial) Protocol. A new HRQL component was developed and approved by the National Cancer Institute and integrated into the study protocol. The P-2 HRQL questionnaire will be given to a sub-sample (n=3000) of the complete STAR cohort (22,000 women); the NSABP application to the Cancer Prevention and Control Protocol Review Committee was approved to give cancer control credits to CCOPS participating in this research. . **Current Status (09/01/01): The HRQL component of the P-2 trial has recruited a total of 2118 participants into the substudy.**

e. **Protocol C-07 - *Trial Comparing 5-Fluorourcil (5-FU) Plus Leucovorin(LV) and Oxaliplatin with 5-FU Plus LV for the Treatment of Patients with Stages II and III Carcinoma of the Colon.*** This study uses the 11 item FACT/GOG-NTX scale in order to obtain the patients' subjective assessment of neurotoxicity attendant upon the administration of Oxaliplatin. **Current Status (09/01/01): The HRQL component of this trial was closed on 7/31/01 following the recruitment of 400 patients to the substudy.**

One new protocol is being developed for inclusion in the P-2 STAR trial:

f. **Protocol STAR-Cog - Effects of Selective Estrogen Receptor Modulators on Cognitive Aging: A Study of Tamoxifen, Raloxifene and Cognition**. This is a direct collaboration with the National Institute of Aging (NIA) and the Woman's Health Initiative Study of Cognitive Aging (WHISCA). The proposed study examines the longitudinal cognitive outcomes in 1800 STAR participants using the same battery of neuropsychological instruments utilized in WHISCA. This protocol is currently awaiting the approval of funding from NIA.

2.2 Testing and implementation of data collection methods to be used in treatment and prevention trials

Operational Procedures to Reduce Missing and Delinquent HRQL Data – Over the past 12 months, procedures intended to reduce missing and delinquent data which were implemented during months 24-36 (i.e., 1999-2000) have been continued, intensified and refined. Specific elements of this strategy include: (1) The use of missing data forms; (2) the inclusion of HRQL questionnaires in delinquency assessments; (3) periodic HRQL training sessions at national meetings; and, (4) the routine notification of study coordinators of scheduled HRQL examinations. Missing and delinquent data continues to be a difficult issue, however, and overall compliance rates for most HRQL studies remain at approximately 70%.

2.3 Analysis of HRQL Data Collected in the NSABP Prevention and Treatment Trials:

a. **Peer Reviewed Papers:**

Day R, Ganz PA, Costantino JC. Tamoxifen and Depression: More Evidence from the NSABP's Breast Cancer Prevention (P-1) Randomized Study. JNCI, 93, 2001 (in press, 7 Nov. 2001 issue). (Appendix 1)

Day R, Quality of life and tamoxifen in breast cancer: a summary of the findings from the NSABP P-1 study. Annals of the New York Academy of Sciences (in press). (Appendix 2)

Land S, Wieand S, Day R, Have T, Costantino J, Lang W, Ganz P. Methodological issues in the analysis of quality of life data in clinical trials: illustrations from the NSABP Breast Cancer Prevention Program. In: M. Mesbah, B. Cole, M Lee (eds.), Statistical Design, Measurement

and Analysis of Health Related Quality of Life. Klewler Academic Publishers (in press). (Appendix 3)

Kiebert G, Wait S, Bernhard J, Bezjak A, Cella D, Day R, Houghton J, Moinpiour C, Scott C, Stephens C. Practice and policy of measuring quality of life and health economics in cancer clinical trials: a survey among cooperative groups. Quality of Life Research 2000; 9(10):1073-80. (Appendix 4)

b. **Submitted papers:**

Day R, Cella D, Ganz PA, Daly MB, Rowland J, Wolter J. Determining the Feasibility and Usefulness of Microelectronic Adherence Monitoring Compared to Pill Counts and Self-Reports in a Large, Multicenter Chemoprevention Trial. Submitted to Controlled Clinical Trials (in revision).

c. **Papers in progress:**

With Dr. Patricia Ganz and Dr. David Cella (mentors): Factor analysis of the P-1 43-item Symptom Checklist data. Initial analyses suggest that most of the variance in baseline SCL scores can be explained by a small number (7 or 8) independent latent variables. The goal of this paper is to simplify the SCL for future prevention studies and assess the stability of these initial latent factors on follow-up in the tamoxifen and placebo arms.

2.4 Refinement and extension of HRQL methods to analyze the data from new treatment and prevention studies

Day, Ganz, and Costantino (2001, Appendix 1) developed and tested methods for assessing depression risk and missing data in the P-1 data. Land, Wieand, Day et al. (2001, Appendix 2) extended and refined missing data techniques in the previous paper.

2.5 Enhancement of minority participation in NSABP trials and the implementation of measures focusing on HRQL-related issues in women of color

This objective is complete.

3. Key Research Accomplishments (Months 37-48)

- Successful implementation of two new NSABP treatment protocols with HRQL components (B-32, B-33)
- Successful completion of recruitment to two ongoing NSABP treatment protocols with HRQL components (B-30, C-07)
- Four peer reviewed articles published or accepted for publication (2 first-authored, 2 co-authored)
- Continued operational supervision of the ongoing HRQL study for the P-2 STAR trial.
- Intensification and refinement of the basic operational strategy to reduce missing and delinquent data in NSABP clinical trials

4. Reportable Outcomes

a. Peer Reviewed Papers:

Day R, Ganz PA, Costantino JC. Tamoxifen and Depression: More Evidence from the NSABP's Breast Cancer Prevention (P-1) Randomized Study. JNCI, 93, 2001 (in press, 7 Nov. 2001 issue). (Appendix 1)

Day R, Quality of life and tamoxifen in breast cancer: a summary of the findings from the NSABP P-1 study. Annals of the New York Academy of Sciences (in press). (Appendix 2)

Land S, Wieand S, Day R, Have T, Costantino J, Lang W, Ganz P. Methodological issues in the analysis of quality of life data in clinical trials: illustrations from the NSABP Breast Cancer Prevention Program. In: M. Mesbah, B. Cole, M Lee (eds.), Statistical Design, Measurement and Analysis of Health Related Quality of Life. Klewler Academic Publishers (in press). (Appendix 3)

Kiebert G, Wait S, Bernhard J, Bezjak A, Cella D, Day R, Houghton J, Moinpiour C, Scott C, Stephens C. Practice and policy of measuring quality of life and health economics in cancer clinical trials: a survey among cooperative groups. Quality of Life Research 2000; 9(10):1073-80. (Appendix 4)

b. Submitted papers:

Day R, Cella D, Ganz PA, Daly MB, Rowland J, Wolter J. Determining the Feasibility and Usefulness of Microelectronic Adherence Monitoring Compared to Pill Counts and Self-Reports in a Large, Multicenter Chemoprevention Trial. Submitted to Controlled Clinical Trials (in revision).

5. Conclusions

There are now a total of 3 ongoing NSABP HRQL studies in the process of data collection (P-2, B-32, B-33). Two studies which were underway at the last reporting date have now been successfully closed to recruitment (B30, C-07). For the past 12 months of the CDA, Dr. Day has had reduced responsibility for day-to-day monitoring of ongoing protocols and has served as a consultant for study development. This has permitted him to devote an increased amount of time in months 37-48 of the grant to the completion of manuscripts and reports. A no cost 12 month extension of the award has been approved enabling Dr. Day to prepare additional analyses and publications on NSABP quality-of-life data.

6. References

None in text.

Appendix 1

Day R, Ganz PA, Costantino JC. Tamoxifen and Depression: More Evidence from the NSABP's Breast Cancer Prevention (P-1) Randomized Study. JNCI, 93, 2001 (in press, 7 Nov. 2001 issue)

ARTICLES

Tamoxifen and Depression: More Evidence From the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Randomized Study

Richard Day, Patricia A. Ganz, Joseph P. Costantino

Background: **Concerns have been raised that tamoxifen may be associated with depression. To investigate this question, we examined the psychological effects of tamoxifen treatment for breast cancer prevention on women at different levels of risk for clinical depression who were enrolled in the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Study.** *Methods:* **A total of 11 064 women were randomly assigned to receive for 5 years daily doses of 20 mg of tamoxifen or placebo in the P-1 study, a multicenter, double-blind, placebo-controlled chemoprevention trial. Each woman was prospectively assessed for depression risk on the basis of medical history items collected at the baseline examination and placed in a high-, medium-, or low-risk group. Every 6 months, for a total of 36 months, the participants were assessed for depressive symptoms by completing the Center for Epidemiological Studies—Depression (CES-D) questionnaire. Scores of 16 or higher were indicative of an episode of affective distress. Differences between the risk groups and treatment arms were analyzed by logistic regression. All statistical tests were two-sided.** *Results:* **Women in the higher risk depression groups were more likely to score 16 or higher on the CES-D (percent follow-up examinations with a score of ≥16: high-risk group = 35.7%, with 95% confidence interval [CI] = 32.5% to 38.9%; medium-risk group = 19.2%, with 95% CI = 18.1% to 20.3%; and low-risk group = 8.7%, with 95% CI = 8.3 to 9.1%) and to have these scores more frequently and for longer periods than women in the lower risk groups. Within each depression risk group, there was no difference in the proportion of women scoring 16 or higher by treatment assignment (tamoxifen versus placebo) (odds ratio = 0.98; 95% CI = 0.93 to 1.02). A *post-hoc* analysis indicated that the lack of a tamoxifen effect was not a result of differential missing data.** *Conclusions:* **Physicians need not be overly concerned that treatment with tamoxifen will increase the risk for or exacerbate existing depression in women. Nevertheless, physicians should continue to screen for and treat or refer potential cases of depression encountered in routine clinical practice. [J Natl Cancer Inst 2001;93:000–00]**

Concern regarding an association between clinical depression and tamoxifen, when used as an adjuvant treatment or preventative agent for breast cancer, has been voiced by a number of investigators *(1–5)* and continues to be discussed in regulatory agencies, such as the U.S. Food and Drug Administration. Furthermore, the *Physician's Desk Reference (6)* lists "depression" as an infrequent adverse reaction to tamoxifen. Although previous studies *(1–5)* used breast cancer patients to address tamoxifen use and depression, the studies had a number of weaknesses, including the lack of a clear definition of depression and a failure to control for the potential confounding effects of illness diagnosis, the side effects of chemotherapy (e.g., premature menopause), or normal aging. Previously, two double-blind, placebo-controlled studies of the effects of tamoxifen in postmenopausal women *(7,8)* found no association of tamoxifen with depression. We believe that some of the concern over the relationship between tamoxifen and depression arises from the idea that, because hormone replacement therapy has positive effects on mood and tamoxifen has antiestrogenic activity *(9–11)*, tamoxifen, therefore, has negative effects on mood.

The completion of the Breast Cancer Prevention (P-1) Study of the National Surgical Adjuvant Breast and Bowel Project (NSABP) provides an opportunity to investigate the association between tamoxifen and depression in greater detail. The P-1 study was a multicenter, double-blind, placebo-controlled chemoprevention trial. The primary objective of the study was to evaluate whether 5 years of tamoxifen therapy would reduce the incidence of invasive breast cancer in women at an increased risk for the disease. The secondary objectives of the study included the assessment of the incidence of ischemic heart disease, bone fractures, and other negative health events, such as depression, that might be associated with tamoxifen therapy. Eligible participants were randomly assigned to receive 20 mg daily of tamoxifen or a placebo for 5 years. Detailed reports on the rationale, planning, design, and clinical outcome of the P-1 study are available elsewhere *(12–16)*.

In our initial publication on the health-related quality of life (HRQL) *(16)* of all subjects in the P-1 study, we did not find a difference between the treatment groups (tamoxifen versus placebo) on the Center for Epidemiological Studies—Depression (CES-D) Scale *(17)* or the SF-36 Mental Health Scale *(18)*. It is known, however, that vulnerability to clinically identifiable forms of depression is not uniformly distributed in the general female population but, instead, clusters in high-risk groups of women *(19)*. This vulnerability to depression may be inherited, suggesting a genetic or familial origin, or it may be related to certain psychological predispositions, such as a low self-esteem, a poor resistance to stress, or a pessimistic view of the world. We

Affiliations of authors: R. Day, J. P. Costantino, National Surgical Adjuvant Breast and Bowel Project Biostatistical Center, Pittsburgh, PA; P. A. Ganz, Jonsson Comprehensive Cancer Center, Los Angeles, CA.

Correspondence to: Richard Day, Ph.D., Department of Biostatistics, Graduate School of Public Health, 130 DeSoto St., University of Pittsburgh, Pittsburgh, PA 15261 (e-mail: day@nsabp.pitt.edu).

See "Notes" following "References."

were concerned that the potential negative effects of tamoxifen for women at high risk for depression may have been masked in our previous analysis *(16)* because of the simultaneous inclusion of a larger group of less vulnerable (i.e., low-risk) participants.

In this study, we investigated the effects of tamoxifen on women at different levels of risk for depression. Specifically, we were interested in whether tamoxifen treatment was associated with the onset or prolongs the length of existing episodes of clinically diagnosable depression in women at high risk for depression.

Patients and Methods

Participant Cohort and HRQL Data

This article covers the baseline and first 36 months of follow-up data (collected at 6-month intervals) on the same 11 064 women used in the initial HRQL report *(16)* from the P-1 study. The P-1 participants ranged in age from 35 years to 79 years (mean ± standard deviation = 53.8 ± 9.2 years), were predominantly white (95.6%), were well educated (≥ some college = 64.9%), and were currently employed (full- or part-time = 64.7%) in a professional or technical field (67.9%). A detailed description of this cohort of participants and the P-1 HRQL instruments was reported previously *(14,16)*. All investigations conducted in the P-1 study were approved by review boards at each institution and were in accord with an assurance filed with and approved by the U.S. Department of Health and Human Services *(12)*. All of the participants provided written informed consent.

Defining Depression

Depressive disorders, as defined by the current psychiatric nomenclature in the *Diagnostic and Statistical Manual of Mental Disorders: DSM- IV* (DSM-IV) *(20)*, come in a variety of forms that differ on the basis of the number, severity, and persistence of symptoms. The majority of clinically diagnosable episodes of depression involve one of three disorders—major depression, dysthymia, or bipolar disorder *(19)*. Major depression involves an illness episode lasting at least 2 weeks that includes mood disturbance (dysphoria) and at least four of the following symptoms: sleep disturbance, change in psychomotor activity, loss of ability to experience pleasure and interest, fatigue, feelings of worthlessness or guilt, difficulty in concentrating, and a preoccupation with death or a wish to die. These symptoms must be associated with a clear impairment in social functioning. Dysthymic disorder or dysthymia is a chronic illness lasting at least 2 years. Dysthymia does not show the same levels of social impairment found in major depression, but it does involve mood disturbance (dysphoria) and a loss of the ability to experience pleasure and interest in usual activities, together with some of the other symptoms used to define major depression. Individuals diagnosed with dysthymia often experience episodes of major depression during their lifetime. DSM-IV distinguishes bipolar disorders from depressive disorders. Bipolar disorders have dramatic clinical manifestations that involve one or more episodes of hypomania during an individual's lifetime alternating with illness episodes that fit the criteria for major depression disorder.

Depression was previously defined by the Research Diagnostic Criteria (RDC) *(21)*, a nonclinical forerunner of the current DSM-IV criteria. The RDC used similar criteria as the DSM-IV to define "major depression" but, unlike the DSM-IV, also included criteria to define "minor depression" (nonpsychotic episodes of illness characterized by a prominent and sustained dysphoria but lacking all of the symptomatic features of major depression). Although important historically, the RDC has been superseded by the DSM-IV.

One of the problems associated with the definition of depression is that, in addition to these diagnosable clinical entities, there are multiple sources of affective distress that may result in short-term or self-limiting expressions of depressive symptoms without meeting the DSM-IV criteria outlined above. The best available data on rates of clinically diagnosable depressive disorders in the U.S. general population come from the National Institute of Mental Health's Epidemiological Catchment Area (ECA) study *(19)*. ECA study investigators found that, even though clinically diagnosable depressive disorders are relatively rare, usually affecting only 5%–6% of the general female population during any 12-month period, the reporting of depressive symptoms is reasonably frequent, with 35.7% of the women in the ECA study *(19)* reporting having experienced a period of dysphoria (feeling sad or blue) lasting at least 2 weeks. These expressions of affective distress, which fail to meet the clinical criteria for major depression, dysthymia, or bipolar illness, are often associated with occurrences such as uncomplicated grief, medical illness and other life events, or chronic difficulties *(22)*. Depressive symptoms may also occur secondary to other psychiatric illnesses (i.e., anxiety disorders or phobias), chronic medical conditions, or substance abuse.

Monitoring Depressive Symptoms in the P-1 Study

The primary instrument used to monitor depressive symptoms in the P-1 study was the CES-D *(17)*. This self-administered questionnaire was designed to be a brief, first-stage screen rather than a clinical diagnostic instrument. The CES-D is composed of 20 items, each of which is scored on a scale of 0–3. Higher scores reflect increased expression of affective distress, and a score of 16 or higher is most often used as the cutoff point for likely cases of clinical depression *(17,23,24)*.

Two problems are associated with the use of the CES-D alone to screen for clinically diagnosable episodes of depression. First, questions on the CES-D inquire only about the past 7 days, collecting little information on the length of time that a symptom has been present. Second, the CES-D collects information only on symptoms and not the degree of social impairment experienced by the respondent. Consequently, scores above the CES-D clinical cutoff point of 16 tend to include a substantial proportion of distressed individuals—perhaps upwards of one half or more—who do not meet the clinical criteria for major depression, dysthymia, or bipolar illness *(24,25)*.

Estimating Depression Risk in P-1 Study Participants

The eligibility criteria for the P-1 study permitted, at the discretion of the local site investigator, the inclusion of women with evidence of clinical depression. Twenty to 22% of the participants scored 16 or higher on the CES-D at least once during any 12-month period of the P-1 study. This percentage exceeds the expected general population rates [5%–6% *(19)*] of clinically diagnosable depressive disorders over a 12-month period by 3.5–4.0 times, indicating that it is necessary to distinguish between clinically diagnosable episodes of depression and depressive symptoms that are secondary to other types of physical and psychiatric illnesses or a consequence of social conditions that produce short-term, self-limiting expressions of affective distress. The preferred means to make such a distinction would be a standardized psychiatric interview, such as the Schedule for Affective Disorders and Schizophrenia—Lifetime Version *(26)* or the Diagnostic Interview Schedule *(19)*. However, in the absence of such an interview, the best single indicator of risk for a future episode of major depression, dysthymia, or bipolar disorder in the P-1 study data is a medical history of treatment for these disorders *(27–30)*.

The ECA study *(19)* found that the mean age at onset for major depressive disorders in the general population was 27 years, with approximately 89% of all first depressive episodes occurring before age 35 years, which was the lower age limit of the participants in the P-1 study. Medical history information, collected on a one-time-only basis as a part of the baseline entry and eligibility assessment of all P-1 study participants, included three self-reported items regarding depression: 1) a medical history of depression, 2) current or previous prescriptions for antidepressant medications, and 3) extended periods (≥12 months) of dysphoric mood (i.e., "depressed or sad most days"). If a participant gave a positive answer to the medical history or the medication question, the interviewer obtained dates of treatment, physicians' names, specific modalities of treatment, and date of last medication dose to assess the consistency and appropriateness of the information provided.

These three medical history items were used in the current study to prospectively estimate each participant's risk of experiencing a clinically diagnosable episode of depression. A simple three-level risk score was determined for each P-1 study participant, depending on whether they endorsed 0 (low risk), 1 or 2 (medium risk), or 3 (high risk) of the medical history items regarding depression in the Entry/Eligibility Form. We hypothesized that women with higher scores on this simple depression risk scale would experience more severe and persistent episodes of affective distress and would be more likely to receive a clinical diagnosis of depression. Moreover, if tamoxifen was associated with the onset and/or prolonged the length of depressive episodes in the high-risk (i.e., more vulnerable) group, it should be apparent from longitudinal differences in the proportion of P-1 study participants in the treatment groups (tamoxifen versus placebo) who scored 16 or higher on the CES-D.

Statistical Analysis

CES-D scores were analyzed as above or below the clinical cutoff of 16 or higher. Binary logistic regression was the primary method of statistical analysis used in this study. Estimated odds ratios (ORs), confidence intervals (CIs), and *P* values are provided for all inferential analyses. Cox regression analysis was used to investigate the effects of treatment and depression risk on the time to the first CES-D with a score of 16 or higher, and Kaplan–Meier curves are provided for these data. When the CES-D data were handled as a continuous variable, nonparametric equivalents to a one-way analysis of variance (i.e., Kruskal–Wallis test) were used because it is unusual for CES-D scores to be normally distributed. Graphic presentations include 95% CIs on observed proportions to provide the reader with visual criteria for the magnitude of potential variation. Reported *P* values are all two-sided and have not been adjusted for multiple statistical comparisons. Instead, we have chosen to focus on consistent patterns of findings rather than on individual statistical tests in forming our conclusions. We also avoided the use of statistical methods for imputation of missing data points in the primary data because the data did not meet the strong assumptions that normally underlie such procedures (e.g., MCAR [i.e., Missing Completely at Random]/MAR [i.e., Missing at Random]). Analyses were carried out with the use of Minitab (Version 13; State College, PA) and Egret (Version 1.0; Cytel Corp., Cambridge, MA).

Au: Give institution at State College, PA

Results

Depression Risk

To determine whether there was an association between depression and tamoxifen treatment in participants of the P-1 study, we first calculated the depression risk score from the frequency of responses to each one of the medical history items (Table 1). The three components of this score were only moderately intercorrelated. The highest correlation occurred between a history of illness and antidepressant medications ($r = .564$; $P<.001$), followed by history of illness and persistent dysphoric mood ($r = .369$; $P<.001$) and medications and dysphoric mood ($r = .269$; $P<.001$). Overall depression risk, measured by the data from this study, was not statistically significantly related to the participants' risk of breast cancer, as measured by the Gail risk model *(12,31)*.

The construct validity of this depression risk score was evaluated, in part, with the use of the social and demographic factors associated with clinically diagnosable depressive disorders in the ECA study *(19)*. Table 2 shows the distribution of the P-1 study participants according to the three-level depression risk scale on seven demographic variables, which approximate those associated with clinically diagnosable depression in the ECA study *(19)*. All of these variables, except education, showed a statistically significant dose–response relationship to the depression risk scores in terms of the direction and intensity of the association.

CES-D Data

Fig. 1, a, shows the proportion of the participants in each depression risk group who scored above the clinical cutoff of 16 or higher on the CES-D Scale at baseline and at each of the follow-up examinations. A consistent, positive dose–response relationship was seen between depression risk, as determined on the basis of the medical history items, and the proportion of participants scoring 16 or higher on the CES-D Scale at each scheduled examination. For each depression risk group, Table 3 shows the mean proportion of follow-up examinations with scores of 16 or higher and the distribution of the maximum and the overall scores on CES-D examinations above the clinical cutpoint. A positive dose–response relationship was also observed between depression risk group and proportion of respondents who scored 16 or higher on sequential CES-D examinations. In the high-risk depression group, for example, 21.2% of the respondents scored 16 or higher on three or more sequential CES-D examinations, compared with 9.7% for the medium-risk group and 3.5% for the low-risk group (data not shown). These findings confirm the expectation that participants in the higher depression risk groups (high>medium>low), on average, tend to experience more persistent and severe episodes of affective distress.

We next analyzed the CES-D data from each depression risk group by treatment group (tamoxifen versus placebo) (Fig. 1, b–d; Table 4). After adjustment for examination and risk group, the results of a logistic regression found that there was a statistically nonsignificant effect for the tamoxifen group compared with the placebo group (OR = 0.98; 95% CI = 0.93 to 1.02; $P = .32$). These analyses indicate that treatment group is not statistically associated with the proportion of women scoring above the CES-D clinical cutoff of 16 or higher in any of the three depression risk groups. Furthermore, after adjustment for depression risk group, an analysis of variance found that

Table 1. Distribution of self-reported risk factors for clinical depressive disorders at baseline examination among participants of the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Study

Risk group (items endorsed)	History of depression	Antidepressant medications	Persistent dysphoria	Risk factor pattern* No.	Risk factor pattern* %	0–3 risk factors† No.	0–3 risk factors† %
Low (0)	No	No	No	7964	72.0	7964	72.0
Medium (1)	No	No	Yes	621	5.6		
	No	Yes	No	668	6.0	1628	14.7
	Yes	No	No	339	3.1		
Medium (2)	No	Yes	Yes	120	1.1		
	Yes	No	Yes	202	1.8	953	8.6
	Yes	Yes	No	631	5.7		
High (3)	Yes	Yes	Yes	519	4.7	519	4.7
Total				11 064	100.0	11 064	100.0

*Depression risk groups were assigned on the basis of the participants' response to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group.

†Number and percent of participants endorsing 0, 1, 2, or 3 depression risk factors.

Table 2. Distribution of NSABP P-1 participants on ECA study social and demographic correlates of clinically diagnosed depressive disorders by depression risk score*

Sociodemographic item	Depression risk score†			Odds ratio‡	95% confidence interval on odds ratio
	Low, %	Medium, %	High, %		
Marital status: divorced/separated	11.1	17.7	23.5	1.63	1.50 to 1.98
Employment status: not working	4.4	7.9	12.2	1.78	1.58 to 2.01
Visited doctor within last 3 mo	71.0	76.4	84.4	1.39	1.28 to 1.51
Hospitalized within last 5 y	42.7	48.6	54.9	1.27	1.19 to 1.36
Age: ≥60 y	29.9	27.4	24.1	0.87	0.81 to 0.94
Education: >high school	66.6	66.7	70.0	1.04	0.97 to 1.12
Income: >median	46.1	37.6	31.3	0.72	0.67 to 0.77

*NSABP P-1 = National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Study; ECA = National Institutes of Mental Health's Epidemiological Catchment Area study [19].

†Depression risk groups were assigned on the basis of the participants' responses to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group.

‡Odds ratios were determined by binary logistic regression; $P<.001$ for all groups compared with referent groups, except for education, where $P=.235$.

there was no difference in the mean individual proportion of follow-up examinations above the clinical cutoff in each treatment arm.

The Kaplan–Meier plot in Fig. 2 shows the relationship between assigned treatment (placebo versus tamoxifen) and depression risk group (high, medium, or low) for the time from randomization until the first CES-D examination with a score exceeding the clinical cutoff of 16 or higher. The results of Cox proportional hazards regression analysis with these data were statistically significant for depression risk group (likelihood

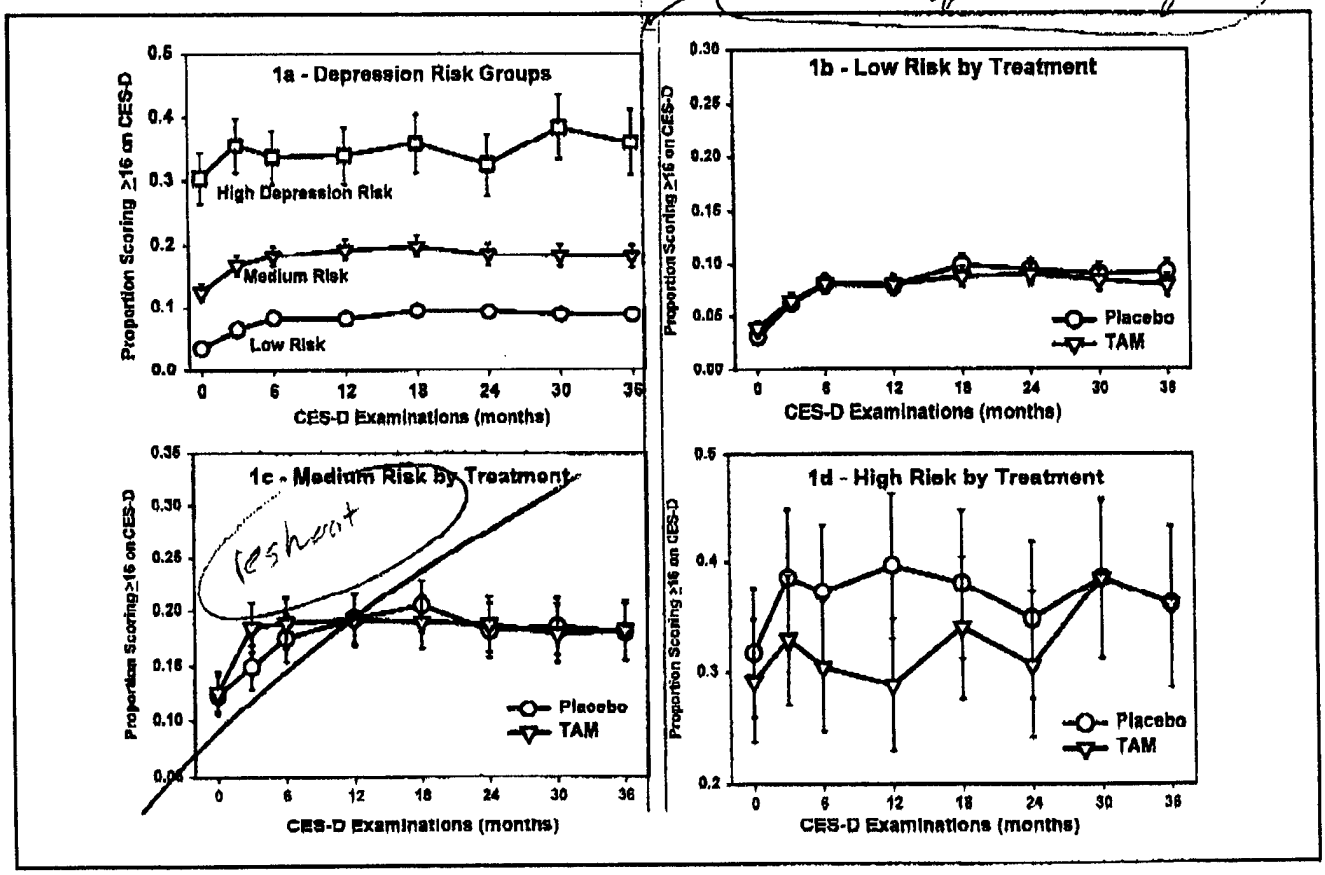


Fig. 1. Proportion of participants in the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Study scoring 16 or higher on the Center for Epidemiological Studies—Depression (CES-D) Scale with 95% confidence intervals by depression risk groups (low, medium, or high) (a) and by depression risk group and treatment assignment (placebo versus tamoxifen [TAM]) (b–d). Depression risk groups were assigned on the basis of the participants' responses to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group.

Table 3. Distribution of Center for Epidemiological Studies—Depression (CES-D) Scale variables for NSABP P-1 participants who scored above the clinical cutoff of 16 or higher by depression risk group*

CES-D variable	Depression risk group†		
	Low	Medium	High
% follow-up examinations in which participants scored ≥16‡			
Mean	0.087	0.192	0.357
95% CI for mean	0.083 to 0.091	0.181 to 0.203	0.325 to 0.389
Maximum score ≥16‡			
Median	22	24	27
Mean	23.97	25.61	28.58
95% CI for mean	23.66 to 24.28	25.16 to 26.06	27.62 to 29.54
All scores ≥16‡			
Median	20	21	22
Mean	21.52	22.49	23.74
95% CI for mean	21.30 to 21.74	22.17 to 22.81	23.10 to 24.38

*The CES-D is a self-administered questionnaire, composed of 20 items, each of which is scored on a scale of 0–3. Higher scores reflect increased expression of affective distress, and a total score of 16 or higher is used as the cutoff point for likely cases of clinical depression *(17,23,24)*. NSABP P-1 = National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Study; CI = confidence interval.

†Depression risk groups were assigned on the basis of the participants' responses to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group.

‡There is a statistically significant difference between all groups (Kruskal–Wallis and analysis of variance; $P<.001$). "Maximum score ≥16" represents the highest single CES-D score ≥16 reported for an individual, whereas "All scores ≥16" summarizes all of the CES-D scores ≥16 reported for an individual.

ratio statistic [LRS] $P<.001$; hazard ratio [HR] = 1.88; 95% CI = 1.74 to 2.05), but they were statistically nonsignificant for both treatment arm effects (LRS $P = .988$; HR = 1.00; 95% CI = 0.92 to 1.09) and interaction effects (LRS $P = .575$; HR = 1.03; 95% CI = 0.92 to 1.16). The proportional hazards assumption for this analysis was confirmed.

Missing Data

We next assessed the association between missing data and depression risk group or sequential CES-D examination (Fig. 3, a). Logistic regression analysis based on the data in Fig. 3, a, indicated that depression risk group (OR = 1.17; 95% CI = 1.13 to 1.21; $P<.001$) and sequential examination (OR = 1.45; 95% CI = 1.44 to 1.46; $P<.001$) were both statistically significantly associated with missing CES-D data. Panels b–d in Fig. 3 show the proportion of participants completing the CES-D by depression risk and treatment groups. Logistic regression analysis by depression risk, controlling for sequential examination, indicates that, compared with placebo treatment, tamoxifen treatment was associated with higher proportions of missing data in the low-risk group (OR = 1.11; 95% CI = 1.06 to 1.16; $P<.001$) and the medium-risk group (OR = 1.12; 95% CI = 1.04 to 1.21; $P<.001$) but not in the high-risk group (OR = 0.99; 95% CI = 0.84 to 1.16; $P = .91$). If tamoxifen-associated depression were the primary cause of these missing data, we would have predicted a positive dose–response increase in the magnitude of the ORs from the lowest to the highest depression risk group.

We noted in our previous report *(15)* that it was difficult to continue to collect quality-of-life data after a participant had gone off treatment. However, participants in the P-1 study were asked about their primary reason for going off treatment, and their responses were recorded on an Off Therapy Form (OTF) that included "depression" as one of 10 specific response categories.

Of the 11 064 participants in this cohort, we collected an OTF for 3539 (80.8%) of 4382 women who missed at least one CES-D examination. The presence of an OTF showed a moderate positive correlation with the total number of missing CES-D examinations ($r = .62$; $P<.001$). The women who completed an OTF accounted for 12 693 (89.7%) of 14 149 missing

Table 4. Comparison (binary logistic regression) of the proportion of NSABP P-1 participants in each treatment group (tamoxifen versus placebo) who scored 16 or higher on the Center for Epidemiological Studies—Depression (CES-D) Scale by depression risk group and sequential examination*

Depression risk group†	Sequential examination							
	Baseline	3 mo	6 mo	12 mo	18 mo	24 mo	30 mo	36 mo
Low								
OR‡	1.22	1.04	1.01	1.02	0.88	0.96	0.93	0.86
95% CI	0.96 to 1.55	0.86 to 1.25	0.85 to 1.19	0.86 to 1.02	0.75 to 1.04	0.80 to 1.13	0.78 to 1.12	0.71 to 1.03
P	.10	.68	.91	0.86	.14	.60	.44	.11
Medium								
OR‡	1.03	1.29	1.10	0.99	0.91	1.04	0.96	1.01
95% CI	0.81 to 1.30	1.04 to 1.60	0.89 to 1.35	0.81 to 1.22	0.74 to 1.13	0.82 to 1.30	0.75 to 1.22	0.79 to 1.29
P	.84	.02	.39	.95	.40	.76	.72	.94
High								
OR‡	0.89	0.78	0.74	0.62	0.84	0.83	1.00	1.00
95% CI	0.61 to 1.30	0.54 to 1.14	0.50 to 1.09	0.41 to 0.92	0.56 to 1.26	0.54 to 1.28	0.65 to 1.54	0.64 to 1.57
P	.54	.21	.13	.02	.40	.40	.99	.99

*The CES-D is a self-administered questionnaire, composed of 20 items, each of which is scored on a scale of 0–3. Higher scores reflect increased expression of affective distress, and a total score of 16 or higher is used as the cutoff point for likely cases of clinical depression *(17,23,24)*. NSABP P-1 = National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Study; OR = odds ratio; CI = confidence interval.

†Depression risk groups were assigned on the basis of the participants' responses to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group.

‡OR >1.0 indicates a greater proportion of women in the tamoxifen group.

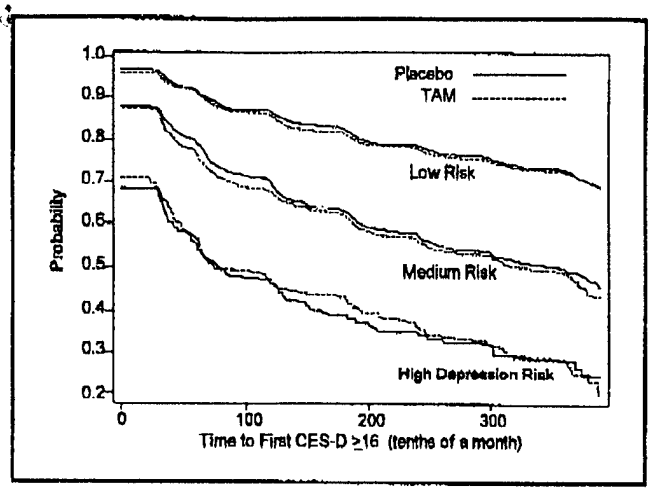


Fig. 2. Kaplan–Meier curves of time from randomization to first score of 16 or higher on the Center for Epidemiological Studies—Depression (CES-D) Scale by depression risk group (low, medium, or high) and treatment assignment (placebo versus tamoxifen [TAM]). Depression risk groups were assigned on the basis of the participants' responses to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group. At 10 months, for the patients who received tamoxifen, in the low-risk group there were 3159 patients at risk of depression (proportion remaining = 0.864; 95% confidence interval [CI] = 0.853 to 0.875); in the medium-risk group there were 799 patients at risk (proportion remaining = 0.685; 95% CI = 0.659 to 0.711); and in the high-risk group there were 123 patients at risk (proportion remaining = 0.488; 95% CI = 0.427 to 0.549). At 30 months, for the patients who received tamoxifen, in the low-risk group there were 2233 patients at risk for depression (proportion remaining = 0.746; 95% CI = 0.732 to 0.760); in the medium-risk group there were 496 patients at risk (proportion remaining = 0.528; 95% CI = 0.499 to 0.557); and in the high-risk group there were 61 patients at risk (proportion remaining = 0.317; 95% CI = 0.258 to 0.376). At 10 months, for the patients who received the placebo, in the low-risk group there were 3190 patients at risk for depression (proportion remaining = 0.870; 95% CI = 0.859 to 0.881); in the medium-risk group there were 863 patients at risk (proportion remaining = 0.713; 95% CI = 0.688 to 0.738); and in the high-risk group there were 108 patients at risk (proportion remaining = 0.475; 95% CI = 0.412 to 0.538). At 30 months, for the patients who received the placebo, in the low-risk group there were 2326 patients at risk for depression (proportion remaining = 0.753; 95% CI = 0.738 to 0.767); in the medium risk group there were 544 patients at risk (proportion remaining = 0.535; 95% CI = 0.506 to 0.563); and in the high-risk group there were 59 patients at risk (proportion remaining = 0.316; 95% CI = 0.254 to 0.377).

CES-D examinations. Only 110 (3.1%) of these 3539 women reported that depression was the primary reason for their going off therapy. The most frequent reasons for going off therapy were nonmedical in nature (1667 women [47.1%]), perceived toxic effects (921 women [26.0%]), and various protocol and nonprotocol medical conditions (841 women [23.8%]).

Table 5 shows the distribution of women who reported that depression was their primary reason for going off treatment by treatment group and depression risk group. An analysis of these data using binary logistic regression found a statistically significant effect for depression risk group (OR = 2.37; 95% CI = 1.83 to 3.07; $P<.001$) and a statistically nonsignificant effect for treatment group (OR = 1.10; 95% CI = 0.75 to 1.62; $P = .63$), indicating that the cases of depression that lead women to quit their assigned treatment did not occur with a greater frequency in those in the tamoxifen arm.

Discussion

Tamoxifen is the most widely prescribed anticancer agent currently in use. It has been proven to be effective against breast cancer as an adjuvant treatment and in a preventative setting *(12,32)*. Given the widespread use of tamoxifen, it is important to fully investigate all of the potential side effects that may be associated with its administration, so that women, together with their physicians, can make an informed decision regarding its potential costs and benefits and its appropriateness for their individual situations.

This study is an extension of our earlier report *(16)* on the HRQL data from the NSABP P-1 study. Previously, we found no evidence for an association between tamoxifen treatment and depression in the overall P-1 study cohort. In this study, we recognized that vulnerability to clinically identifiable depressive disorders is not randomly distributed in the general female population and that the effects of tamoxifen on susceptible women in the P-1 study may have previously gone undetected.

Our initial problem was the *a priori* identification of subgroups of women with a potential clinical susceptibility for depression. Because the self-administered depression-screening form (CES-D) used in the P-1 study provides information on short-term symptoms of affective distress and is not intended for use as a diagnostic instrument *(17)*, we incorporated the participants' self-reported medical history of depression, use of prescription antidepressant medications, and experience of extended periods (>12 months) of dysphoric mood to assign clinical risk. On the basis of these data, women were prospectively assigned to one of three depression risk groups. We hypothesized that the higher a woman's depression risk group, the greater the likelihood that she would experience a clinically diagnosable episode of depression.

The P-1 study staff were trained to check the consistency and appropriateness of the self-reported data about prior treatment for depression and the use of antidepressant medications as a routine part of the medical screening procedure carried out during entry/eligibility interview. These procedures were designed specifically to minimize false-positive classification errors. However, there was little that the interviewer could do to detect false-negative classification errors in which a potential participant did not, for whatever reason, report the requested screening information. The overall effect of this inability to control for false-negative classification errors for the current study was to create a potential misclassification bias in which women at increased risk for depression may have been placed, at an unknown rate, in one of the lower risk groups. Although less than ideal, the effect of this bias is conservative in nature, operating to maintain the comparative validity of the most important high-risk depression group.

We found a statistically significant dose–response relationship between the level of the depression risk group (high> medium>low) and the proportion of the women in each depression risk group who scored above the clinical cutoff of 16 or higher on the CES-D at baseline and at every follow-up interview. In addition, women in the higher risk groups (high> medium>low) scored above the clinical cutoff on a greater proportion of their follow-up interviews and, on average, had higher maximum CES-D scores. Together, these data suggest that there was a dose–response effect, in which women in the higher depression risk groups (high>medium>low) were more likely to

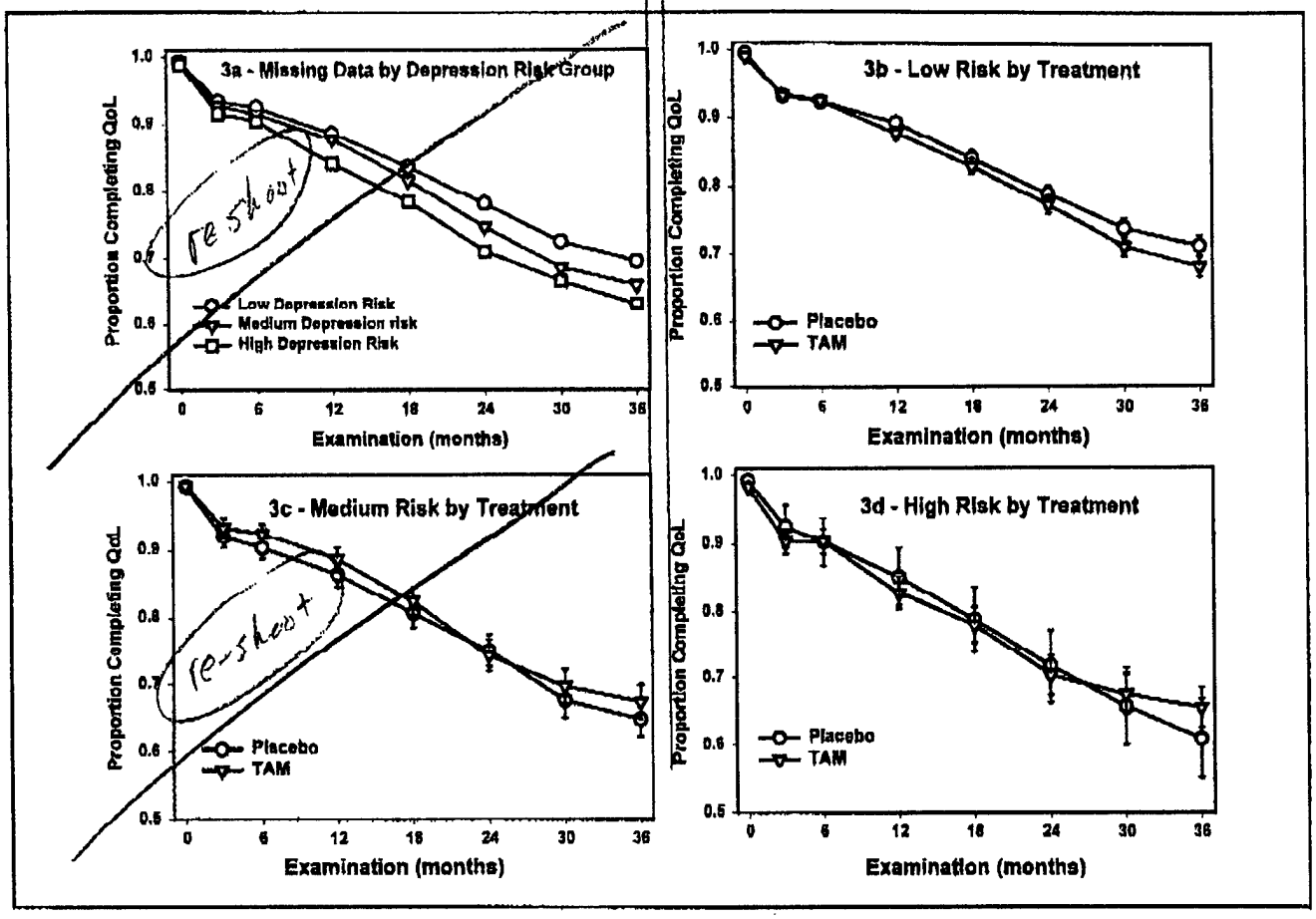


Fig. 3. Proportion of participants in the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Study completing the health-related quality-of-life questionnaire by depression risk groups (low, medium, or high) (**a**) and by depression risk group and treatment assignment (placebo versus tamoxifen [TAM]) with 95% confidence intervals (**b–d**). Depression risk groups were assigned on the basis of the participants' responses to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group.

Table 5. Reasons cited for going off treatment by depression risk* and treatment group

	Low risk		Medium risk		High risk		
Reasons cited for going off treatment	Placebo	Tamoxifen	Placebo	Tamoxifen	Placebo	Tamoxifen	Overall
Depression (No. of participants)	20	27	21	24	9	9	110
Other reasons (No. of participants)	1130	1275	416	431	83	94	3429
Depression as % of all off-treatment reasons	1.7	2.1	4.8	5.3	9.8	8.7	3.1

*Depression risk groups were assigned on the basis of the participants' responses to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group.

experience clinically significant episodes of affective distress and that these episodes, on average, were more persistent and more severe than the episodes in the lower risk groups. Finally, we found that the distribution of social and demographic correlates (i.e., age, marital and employment status, educational level, and use of medical services) across the three depression risk groups defined in this study followed the same general patterns of risk previously identified in the ECA study of depression among the general population *(21)*. All of the above findings serve to support the validity of the risk assignments used in our study.

The primary test of our research question involved stratifying each depression risk group by treatment assignment (tamoxifen versus placebo) and comparing the corresponding proportions of women at each follow-up interview who scored above the clinical cutoff of 16 or higher on the CES-D. We found no effect of tamoxifen for any of the three depression risk groups.

Besides the lack of a positive association between tamoxifen use and depression, there are at least two possible alternative explanations for our negative findings: lack of statistical power and missing data. We carried out a *post-hoc* effect size analysis to determine the size of the difference between the treatment

arms that might have been detected. For our highest risk depression group (n = 519), we had an 80% chance of detecting at least a 37% (OR ≥1.37) increase between the two study arms in the proportions of women scoring above the CES-D clinical cutoff of 16 or higher at any single examination point. When a repeated measures design was used, we had sufficient power to detect a mean increase of 24% (OR ≥1.24) in the proportion of women in either arm scoring above the CES-D clinical cutoff *(33,34)*. We considered these to be acceptable levels of statistical power for the identification of clinically significant treatment effects in our high-risk depression group. The detectable ORs were, of course, even smaller for the low- and medium-risk depression groups.

We also assessed the contribution of missing data to explain the negative association between tamoxifen and depression in the P-1 study. An initial analysis showed that assigned depression risk was statistically significantly associated with missing data rates over the course of the study. If a tamoxifen-associated depression was the primary cause of these rates, we would have predicted that the tamoxifen treatment group in the higher depression risk groups would show a progressively greater differential off-treatment rate than the placebo group. This expectation was not confirmed by our data for the high-risk depression group.

In addition, we also examined the reasons given for going off the assigned treatment. There was a strong statistical association in the P-1 study between stopping assigned treatment and missing HRQL data *(16)*. An analysis of the reasons for going off treatment in 81% of the women with missing HRQL data resulted in the following observations: (*a*) Depression was cited as a relatively infrequent reason for going off treatment; (*b*) the higher the depression risk group, the greater the likelihood that depression was cited as the reason for going off treatment; and (*c*) within each depression risk group, depression was cited as the reason for going off treatment by similar proportions of women, regardless of treatment assignment. A separate report *(35)* has implemented a sensitivity analysis on these data with equally negative results. The findings in our report together with this sensitivity analysis indicates that there are no clear patterns in the missing data that serve to undermine the conclusions drawn from our primary analysis.

The results of our analysis strengthen our previous conclusion regarding lack of evidence for an association between tamoxifen use and depression in the P-1 study data by provisionally extending our findings to subgroups of women at a high risk for clinically identifiable episodes of depression. Clinically, these findings have two major implications. First, the evidence from NSABP's P-1 study does not lend support to the idea that tamoxifen should be considered to be a causal risk factor for the onset of depressive symptoms and/or the prolongation of depressive episodes that occur among treated women. Second, the findings of this study suggest that physicians need not automatically disqualify women as candidates for tamoxifen treatment simply because they report a history of depressive symptoms or prior treatment for a depressive disorder. Nevertheless, it is still essential that physicians carefully screen for affective disorders and treat or refer potential cases of depression encountered in routine clinical practice.

Finally, there are two important limitations on these conclusions that require discussion, one statistical and the other methodological. Statistically, it was the large size of the P-1 study that permitted us to identify and carry out stratified analyses of groups of women with a differential risk for depression. However, we also noted that there were limits on our statistical power to detect an increase in the proportion of women reporting clinically significant levels of depressive symptoms on the CES-D, particularly in the high-risk depression group. For this reason, we cannot absolutely exclude the possibility that there may be rare cases in which women react negatively to tamoxifen treatment with potentially life-threatening depressions. Here, it is useful to recall that data on neuro-mood toxic effects were collected for P-1 study participants and periodically reviewed as part of the routine safety-monitoring procedures. Over the full course of the P-1 study, there were a total of three women who committed suicide, one woman from the placebo-treated group and two women from the tamoxifen-treated group, and there were no statistically significant differences in the distribution of women reporting suicidal ideation across the two trial arms.

The methodological limitations of this article (i.e., the lack of standardized psychiatric diagnoses and missing HRQL data) are primarily due to the fact that the goals of this study were secondary to the main clinical objectives that determined the design of the P-1 study. A more definitive analysis would require additional data from a potentially smaller, yet more focused study, in which an investigation of the relationship between clinical depression and tamoxifen treatment was the primary scientific objective. Such a study would have to have the following minimum features: (*a*) a double-blind, placebo-controlled, randomized design; (*b*) participants who are at high risk for breast cancer, rather than breast cancer patients (to avoid potential confounding due to clinical diagnosis and treatment); (*c*) participants who are stratified on a reliable measure of risk for affective disorder (e.g., lifetime diagnosis, Schedule for Affective Disorders and Schizophrenia—Lifetime Version); (*d*) periodic administration, in whole or in part, of a standardized psychiatric diagnostic instrument (e.g., Diagnostic Interview Schedule) by a trained interviewer; and (*e*) continued collection of the psychiatric interview data even if the participant goes off the assigned treatment for any reason, except death or consent withdrawal. Whether the additional information obtained from such a study would justify the time and the expense involved in its collection is a problematic question that is beyond the scope of this article.

References

(1) Cathcart CK, Jones SE, Pumroy CS, Peters GN, Knox SM, Cheek JH. Clinical recognition and management of depression in node negative breast cancer patients treated with tamoxifen. Breast Cancer Res Treat 1993;27: 277–81.

(2) Shariff S, Cumming CE, Lees A, Handman M, Cumming DC. Mood disorder in women with early breast cancer taking tamoxifen, an estradiol receptor antagonist. An expected or unexpected effect? Ann N Y Acad Sci 1995;761:365–8.

(3) Anelli TF, Anelli A, Tran KN, Lebwohl DE, Borgen PI. Tamoxifen administration is associated with a high rate of treatment-limiting symptoms in male breast cancer patients. Cancer 1994;74:74–7.

(4) Pluss JL, DiBella NJ. Reversible central nervous system dysfunction due to tamoxifen in a patient with breast cancer. Ann Intern Med 1984;101:652.

(5) Duffy LS, Greenberg DB, Younger J, Ferraro MG. Iatrogenic acute estrogen deficiency and psychiatric syndromes in breast cancer patients. Psychosomatics 1999;40:304–8.

(6) Physician's Desk Reference. 53rd ed. Oradell (NJ): Medical Economics Co.; 1999.

(7) Love RR, Cameron L, Connell BL, Leventhal H. Symptoms associated with tamoxifen treatment in postmenopausal women. Arch Intern Med 1991;151:1842–7.
(8) Fallowfield L, Fleissig A, Edwards R, West A, Powles TJ, Howell A, et al. Tamoxifen for the prevention of breast cancer: psychosocial impact on women participating in two randomized controlled trials. J Clin Oncol 2001;19:1885–92.
(9) Holbreich U. Role of estrogen in postmenopausal depression. Neurology 1997;48(5 Suppl 7):S16–9.
(10) Gregoire AJ, Kumar R, Everitt B, Henderson AF, Studd JW. Transdermal oestrogen for treatment of severe postnatal depression. Lancet 1996;347: 930–3.
(11) Murray D. Oestrogen and postnatal depression. Lancet 1996;347:918–9.
(12) Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, et al. Tamoxifen for the prevention of breast cancer: report from the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Natl Cancer Inst 1998;90:1371–88.
(13) Fisher B, Costantino J. Highlights of the NSABP Breast Cancer Prevention Trial. Cancer Control 1997;4:78–86.
(14) Ganz PA, Day R, Ware JE Jr, Redmond C, Fisher B. Base-line quality-of-life assessment in the National Surgical Adjuvant Breast and Bowel Project Breast Cancer Prevention Trial. J Natl Cancer Inst 1995;87:1372–82.
(15) Ganz PA, Day R, Costantino JP. Compliance with quality of life data collection in the National Surgical Adjuvant Breast and Bowel Project (NSABP) Breast Cancer Prevention Trial. Stat Med 1998;17:613–22.
(16) Day R, Ganz PA, Costantino JP, Cronin WM, Wickerham DL, Fisher B. Health-related quality of life and tamoxifen in breast cancer prevention: a report from the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Clin Oncol 1999;17:2659–69.
(17) Radloff LS. The CES-D scale: a self-report depression scale for use in the general population. Applied Psychol Measurement 1977;1:385–401.
(18) Stewart AL, Ware JE. Measuring functioning and well-being: the medical outcomes approach. Durham (NC): Duke University Press; 1992.
(19) Weissman MM, Bruce ML, Leaf PJ, Florio LP, Holzer C. Affective disorders. In: Robins LN, Regier DA, editors. Psychiatric disorders in America: the Epidemiological Catchment Area study. New York (NY): The Free Press; 1991.
(20) Diagnostic and statistical manual of mental disorders: DSM-IV. Washington (DC): American Psychiatric Association; 2000.
(21) Spitzer RL, Endicott J, Robbins E. Research Diagnostic Criteria. New York (NY): Biometrics Research Division, New York State Psychiatric Institute; 1978.
(22) Dohrenwend BP, Shrout PE, Egin G, Mendelsohn FS. Nonspecific psychological distress and other dimensions of psychopathology. Measures for use in the general population. Arch Gen Psychiatry 1980;37:1229–36.
(23) Myers JK, Weissman MM. Use of a self-report symptom scale to detect depression in a community sample. Am J Psychiatry 1980;137:1081–4.
(24) Lin N, Dean A, Ensel WM. Social support, life events, and depression. New York (NY): Academic Press; 1986.
(25) Boyd JH, Weissman MM, Thompson WD, Myers JK. Screening for depression in a community sample. Understanding the discrepancies between depression symptom and diagnostic scales. Arch Gen Psychiatry 1982;39: 1195–200.
(26) Spitzer RL, Endicott J. Schedule for Affective Disorders and Schizophrenia. New York (NY): Biometrics Research Division, New York State Psychiatric Institute; 1978.
(27) Kaelber CT, Moul DE, Farmer ME. Epidemiology of depression. In: Beckham EE, Leber WR, editors. Handbook of depression. 2nd ed. New York (NY): Guilford Press; 1995. p. 3–35.
(28) Keller MB, Lavori PW, Rice J, Coryell W, Hirschfeld RM. The persistent risk of chronicity in recurrent episodes of nonbipolar major depressive disorder: a prospective follow-up. Am J Psychiatry 1986;143:24–8.
(29) Keller MB, Lavori PW, Lewis CE, Klerman GL. Predictors of relapse in major depressive disorder. JAMA 1983;250:3299–304.
(30) Leaf PJ, Livingston MM, Tischler GL, Weissman MM, Holzer CE 3rd, Myers JK. Contact with health professionals for the treatment of psychiatric and emotional problems. Med Care 1985;23:1322–37.
(31) Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst 1999;91:1541–8.
(32) Fisher B, Costantino J, Redmond C, Poisson R, Bowman D, Couture J, et al. A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors. N Engl J Med 1989;320:479–84.
(33) Fleiss JH. Statistical methods for rates and proportions. 2nd ed. New York (NY): John Wiley & Sons; 1981.
(34) Brown H, Prescott R. Applied mixed models in medicine. New York (NY): John Wiley & Sons; 1999.
(35) Land S, Wieand S, Day R, Ten Have T, Costantino JP, Lang W, et al. Methodological issues in the analysis of quality of life data in clinical trials: illustrations from the National Surgical Adjuvant Breast and Bowel Project (NSABP) Breast Cancer Prevention Trial. In: Mesbah M, Cole B, Lee M, editors. Statistical design, measurement and analysis of health-related quality of life. New York (NY): Kluwer Academic Publishers. In press 2001.

NOTES

Supported by Public Health Service grant NCI-U10CA37377/69974 from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services; by career development award DAMD17-97-1-7058 from the Department of Defense (to R. Day); and in part by an American Cancer Society Clinical Research Professorship (to P. A. Ganz).

We thank Samuel Wieand, Ph.D, Stephanie Land, Ph.D., and Ms. Sheela Goshal of the National Surgical Adjuvant Breast and Bowel Project (NSABP) Biostatistical Center and D. Lawrence Wickerham, M.D., of the NSABP Operations Center for their help in the preparation of this article.

Manuscript received February 20, 2001; revised August 15, 2001; accepted August 31, 2001.

Tamoxifen and Depression:

More Evidence from the National Surgical Adjuvant Breast and Bowel Project Breast Cancer Prevention (P-1) Randomized Study

Richard Day, Patricia A. Ganz, Joseph P. Costantino

Key Words: Breast Cancer, Depression, Quality-of-Life, Tamoxifen

Affiliations of authors: Richard Day, Joseph P Costantino, National Surgical Adjuvant Breast and Bowel Project (NSABP) Biostatistical Center, Pittsburgh, PA, Patricia A. Ganz, Jonsson Comprehensive Cancer Center, Los Angeles, CA. *Correspondence to:* Richard Day, Ph.D., Department of Biostatistics, Graduate School of Public Health, 130 DeSoto Street, University of Pittsburgh, Pittsburgh, PA 15261. E-mail: day@nsabp.pitt.edu.

ABSTRACT

Background: We examined the psychological effects of tamoxifen treatment for the breast cancer prevention on women at different levels of risk for clinical depression to determine whether tamoxifen treatment is associated with the onset of and/or prolongs the length of existing episodes of clinical depression in psychologically vulnerable individuals from the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) study. *Methods:* We report on the baseline and first 36 months of follow-up data for 11,064 women, randomly assigned to daily doses of 20 mg of tamoxifen or placebo in the P-1 study, which is a multi-center, double-blind, placebo-controlled chemoprevention trial. Each woman was prospectively assessed for depression risk on the basis of medical history items collected at the baseline examination and placed in a high-, medium- or low-risk group. Every 6 months, the participants were assessed for depressive symptoms by completing the Center for Epidemiological Studies-Depression (CES-D) questionnaire. Scores of ≥16 were indicative of an episode of affective distress. Differences between the risk groups and treatment arms were assessed by logistic regression. All statistical tests were two-sided. *Results:* Women in the high-risk depression group were more likely to score ≥16 on the CES-D, more frequently and for longer time intervals than women in the lower risk groups. Within the depression risk groups, there was no difference in the proportion of women scoring ≥16 by treatment assignment (tamoxifen versus placebo) (OR = 0.98; 95% CI = 0.93 to 1.02). The lack of a tamoxifen effect was not a result of differential missing data. *Conclusions:* Physicians need not be overly concerned that treatment with tamoxifen will increase the risk for or exacerbate existing depression in women.

Nevertheless, physicians should screen for and treat or refer potential cases of depression encountered in routine clinical practice.

INTRODUCTION

Concern regarding an association between clinical depression and tamoxifen, when used as an adjuvant treatment or preventative agent for breast cancer, has been voiced by a number of investigators *(1-5)* and continues to be discussed in regulatory agencies, such as the U.S. Food and Drug Administration. Furthermore, the Physician's Desk Reference *(6)* lists "depression" as an infrequent adverse reaction to tamoxifen. Although prior studies *(1-5)* used breast cancer patients to address tamoxifen use and depression, the studies had a number of weaknesses, including the lack of a clear definition of depression and a failure to control for the potential confounding effects of illness diagnosis, the side-effects of chemotherapy (e.g., premature menopause), or normal aging. Previously, two double-blind, placebo-controlled studies of the effects of tamoxifen in post-menopausal women *(7-8)* found no association with depression. We believe that some of the concern over the relationship between tamoxifen and depression arises from the idea that because hormone replacement therapy has positive effects on mood and tamoxifen has anti-estrogenic activity *(9-11),* tamoxifen, therefore, has negative effects on mood.

The completion of the National Surgical Adjuvant Breast and Bowel Project (NSABP) Breast Cancer Prevention Trial (P-1) provides an opportunity to investigate the association between tamoxifen and depression in greater detail. The P-1 study was a multi-center, double-blind, placebo-controlled chemoprevention trial. The primary objective of the study was to evaluate whether 5 years of tamoxifen therapy would

reduce the incidence of invasive breast cancer in women at an increased risk for the disease. The secondary objectives included assessing the incidence of ischemic heart disease, bone fractures, and other negative health events, such as depression, that might be associated with tamoxifen therapy. Eligible participants were randomly assigned to receive 20 mg daily of tamoxifen or a placebo for 5 years. Detailed reports on the rationale, planning, design, and clinical outcome of the P-1 study are available elsewhere *(12-16)*.

In our initial health-related quality of life (HRQL) publication *(16)* of all subjects in the P-1 study, we did not find a difference between the treatment groups (tamoxifen versus placebo) on the Center for Epidemiological Studies-Depression (CES-D) scale *(21)* or the SF-36 Mental Health scale *(35)*. It is known, however, that vulnerability to clinically identifiable forms of depression is not uniformly distributed in the general female population, but, instead, clusters in high-risk groups of women *(17)*. This vulnerability to depression may be inherited, suggesting a genetic or familial origin, or it may be related to certain psychological predispositions, such as a low self-esteem, a poor resistance to stress, or a pessimistic view of the world. We were concerned that the potential negative effects of tamoxifen for women at high-risk for depression may have been masked in our previously analysis *(16)* because of the simultaneous inclusion of a larger group of less vulnerable (i.e., low-risk) participants. In this study, we investigated the effects of tamoxifen for women at different levels of risk for depression. Specifically, we were interested in whether tamoxifen treatment was associated with the onset or prolongs the length of existing episodes of clinically diagnosable depression in women at high-risk for depression.

PATIENTS AND METHODS

Participant cohort and HRQL data

This article covers the baseline and first 36 months of follow-up data on the same 11,064 women used in the initial HRQL report *(16)* from the P-1 study. The P-1 participants' ranged in age from 35-79 years (mean = 53.8 ± 9.2 years), were predominately white (95.6%), well educated (≥ some college = 64.9%), and currently employed (full or part-time = 64.7%) in a professional or technical field (67.9%). A detailed description of this cohort of participants and the P-1 HRQL instruments are available in previous publications *(14,16)*. All investigations conducted in the P-1 study were approved by review boards at each institution and were in accord with an assurance filed with and approved by the U.S. Department of Health and Human Services *(12)*. All participants provided written, informed consent.

Defining Depression

Depressive disorders, as defined by the current psychiatric nomenclature in the Diagnostic and Statistical Manual of Mental Disorders IV (DSM-IV, *18*), come in a variety of forms that differ on the basis of the number, severity, and persistence of symptoms. The majority of clinically diagnosable episodes of depression involve one of three disorders - major depression, dysthymia, or a bipolar illness (17). Major depression involves an illness episode lasting at least 2 weeks that includes mood disturbance (dysphoria) and at least four of the following symptoms: sleep disturbance, change in psychomotor activity, loss of ability to experience pleasure and interest, fatigue, feelings of worthlessness or guilt, difficulty in concentrating, and a preoccupation with death or a wish to die. These symptoms must be associated with a

clear impairment in social functioning. Dysthymic disorder or dysthymia is a chronic illness lasting at least 2 years. Dysthymia does not show the same levels of social impairment found in major depression, but does involve mood disturbance (dysphoria) and a loss of the ability to experience pleasure and interest in usual activities, together with some of the other symptoms used to define major depression. Individuals diagnosed with dysthymia often experience episodes of major depression during their lifetime. DSM-IV distinguishes bipolar disorders from depressive disorders. Bipolar disorders have dramatic clinical manifestations that involve one or more episodes of hypomania during an individual's lifetime alternating with illness episodes that fit the criteria for major depression disorder.

Depression was previously defined by the Research Diagnostic Criteria (RDC, *19*), a non-clinical forerunner of the current DSM-IV criteria. The RDC used similar criteria as the DSM-IV to define "major depression" and included a diagnosis of "minor depression" (nonpsychotic episodes of illness characterized by a prominent and sustained dysphoria, but lacking all the symptomatic features of major depression) that was not included in the DSM-IV. Although important historically, the RDC has been superseded by the DSM-IV.

One of the problems associated with the definition of depression is that in addition to these diagnosable clinical entities, there are multiple sources of affective distress that may result in short-term or self-limiting expressions of depressive symptoms without meeting the DSM IV criteria outlined above. The best available data on rates of clinically diagnosable depressive disorders in the United States general population come from the National Institute of Mental Health's Epidemiological Catchment Area Study (ECA, *17*). ECA investigators found that, even though clinically

diagnosable depressive disorders are relatively rare, usually affecting only 5-6% of the general female population during any 12-month period, the reporting of depressive symptoms is reasonably frequent, with 35.7% of the women in the ECA study *(17)* reporting having experienced a period of dysphoria (feeling sad or blue) lasting at least 2 weeks. These expressions of affective distress which fail to meet the clinical criteria for major depression, dysthymia, or bipolar illness are often associated with occurrences such as uncomplicated grief, medical illness and other life events, or chronic difficulties *(20)*. Depressive symptoms may also occur secondary to other psychiatric illnesses (i.e., anxiety disorders or phobias), chronic medical conditions, or substance abuse.

Monitoring Depressive Symptoms in the P-1 Study

The primary instrument used to monitor depressive symptoms in the P-1 study was the CES-D *(21)*. This self-administered questionnaire was designed to be a brief, first stage screen rather than a clinical diagnostic instrument. The CES-D is composed of 20 items, each of which is scored on a scale of 0-3. Higher scores reflect increased expression of affective distress and a score of ≥16 is most often used as the cut-off point for likely cases of clinical depression *(21-23)*.

There are two problems associated with use of the CES-D alone to screen for clinically diagnosable episodes of depression. First, questions on the CES-D only inquire about the past 7 days, collecting little information on the length of time that a symptom has been present. Second, the CES-D only collects information on symptoms and not the degree of social impairment experienced by the respondent. Consequently, scores above the CES-D clinical cut-off point of 16 tend to include a substantial

proportion of distressed individuals - perhaps, upwards of one-half or more - who do not meet the clinical criteria for major depression, dysthymia, or bipolar illness (23,24).

Estimating Depression Risk in P-1 Participants

The P-1 eligibility criteria permitted, at the discretion of the local site investigator, the inclusion of women with evidence of clinical depression. Twenty to 22% of the participants scored ≥16 on the CES-D at least once during any 12-month period of the P-1 study. This figure exceeds the expected general population rates (5-6%, [*17*]) of clinically diagnosable depressive disorders over a 12-month period by 3.5-4.0 times, indicating that it is necessary to distinguish between clinically diagnosable episodes of depression and depressive symptoms that are secondary to other types of physical and psychiatric illnesses or a consequence of social conditions that produce short term, self-limiting expressions of affective distress. The preferred means to make such a distinction would be a standardized psychiatric interview, such as the Schedule for Affective Disorders and Schizophrenia - Lifetime Version *(25)* or the Diagnostic Interview Schedule *(17)*. However, in the absence of such an interview, the best single indicator of risk for a future episode of major depression, dysthymia, or bipolar disorder in the P-1 data is a prior medical history of treatment for these disorders *(26-29.*

The ECA study *(17)* found that the mean age at onset for major depressive disorders in the general population was 27 years, with approximately 89% of all first depressive episodes occurring before age 35 years, which was the lower age limit of the participants in the P-1 study. Medical history information, collected on a one-time only basis as a part of the baseline entry and eligibility assessment of all P-1 participants, included three self-reported items regarding depression: (a) a prior medical history of depression; (b) current or previous prescriptions for anti-depressant medications; and

(c) extended periods (12 or more months) of dysphoric mood (i.e., "depressed or sad most days"). If a participant gave a positive answer to either the medical history or the medication question, the interviewer obtained dates of treatment, physicians' names, specific modalities of treatment, and date of last medication dose to assess the consistency and appropriateness of the information provided.

These three medical history items were used in the current study to prospectively estimate each participant's risk of experiencing a clinically diagnosable episode of depression. A simple three level risk score was determined for each P-1 participant depending on whether they endorsed 0 (low-risk), 1 or 2 (medium-risk), or 3 (high-risk) of the medical history items regarding depression in the Entry/Eligibility Form. We hypothesized that women with higher scores on this simple depression risk scale would experience more severe and persistent episodes of affective distress, and would be more likely to receive a clinical diagnosis of depression. Moreover, if tamoxifen was associated with the onset and/or prolonged the length of depressive episodes in the high-risk (i.e., more vulnerable) group, it should be apparent from longitudinal differences in the proportion of P-1 participants in the treatment groups (tamoxifen versus placebo) who scored ≥16 on the CES-D.

Statistical Analysis

CES-D scores were analyzed as above or below the clinical cut-off of ≥16. Binary logistic regression was the primary method of statistical analysis used in this study. Estimated risk ratios (ORs), confidence intervals (CIs), and *P*-values are provided for all inferential analyses. Cox regression analysis was used to investigate the effects of treatment and depression risk on the time to the first CES-D with a score ≥16 and Kaplan-Meier plots are provided for these data. When the CES-D data were handled as

a continuous variable, nonparametic equivalents to a one-way ANOVA (i.e., Kruskal-Wallis test) were used because it is unusual for CES-D scores to be normally distributed. Graphic presentations include 95% CIs on observed proportions to provide the reader with visual criteria for the magnitude of potential variation. Reported *P*-values are all two-sided and have not be adjusted for multiple statistical comparisons. Instead, we have chosen to focus on consistent patterns of findings, rather than individual statistical tests in forming our conclusions. We also avoided the use of statistical methods for imputation of missing data points in the primary data because the data did not meet the strong assumptions that normally underlie such procedures (e.g., MCAR/MAR). Analyses were carried out using Minitab (Ver. 13) and Egret (Ver. 1.0).

RESULTS

Depression risk

To determine whether there was an association between depression and tamoxifen treatment in participants of the P-1 study, we first calculated the depression risk score from the frequency of responses to each one of the medical history items (Table 1). The three components of this score were only moderately intercorrelated. The highest correlation occurred between a prior history of illness and anti-depressant medications ($r = 0.564$, $P < .001$), followed by prior history and persistent dysphoric mood ($r = 0.369$, $P < .001$), and medications and dysphoric mood ($r = 0.269$, $P < .001$). Overall depression risk, measured by the data from this study, was not statistically significantly related to the participants' risk of breast cancer, as measured by the Gail risk model *(12, 30)*.

The construct validity of this depression risk score was evaluated, in part, with the use of the social and demographic factors associated with clinically diagnosable depressive disorders in the ECA study *(17)*. Table 2 shows the distribution of P-1 participants according to the three level depression risk scale on seven demographic variables, which approximate those associated with clinically diagnosable depression in the ECA study *(17)*. All of these variables, except education, showed a statistically significant dose-response relationship to the depression risk scores in terms of the direction and intensity of the association.

CES-D data

Figure 1a shows the proportion of the participants in each depression risk group who scored above the clinical cut-off of ≥16 on the CES-D at baseline and at each of the follow-up examinations. A consistent, positive dose-response relationship exists between depression risk, as determined on the basis of the of medical history items, and the proportion of participants scoring ≥16 on the CES-D at each scheduled examination. For each depression risk group, Table 3 shows the mean proportion of follow-up examinations with scores ≥16, and the distribution of the maximum and the overall scores on CES-D examinations above the clinical cut-point. A positive dose-response relationship was also observed between depression risk group and proportion of respondents who scored ≥16 on sequential CES-D examinations. In the high-risk depression group, for example, 21.2% of the respondents scored ≥16 on three or more sequential CES-D examinations, compared with 9.7% for the medium-risk and 3.5% for the low-risk groups. These findings confirm the expectation that participants in the

higher depression risk groups (low>medium>high), on average, tend to experience more persistent and severe episodes of affective distress.

We next analyzed the CES-D data from each depression risk group by treatment group (tamoxifen versus placebo) (Figures 1b-d and Table 4). After adjusting for examination and risk group, the results of a logistic regression found that there was a statistically non-significant effect for the tamoxifen group compared with the placebo group (OR=0.98; 95% CI = 0.93 to 1.02; P = .32). These analyses indicate that treatment group is not statistically associated with the proportion of women scoring above the CES-D clinical cut-off of ≥16 in any of the three depression risk groups. Furthermore, after adjusting for depression risk group, an analysis of variance found that there was no difference in the mean individual proportion of follow-up examinations above the clinical cut-off in each treatment arm.

The Kaplan-Meier plot in Figure 2 shows the relationship between assigned treatment (placebo/tamoxifen) and depression risk group (high/medium/low) for the time from randomization until the first CES-D examination with a score exceeding the clinical cut-off of ≥16. The results of Cox proportional hazards regression analysis with these data were statistically significant for depression risk group (likelihood ratio statistic P < .001; hazard ratio=1.88 95% CI: 2.05-1.74), but statistically nonsignificant for both treatment arm (LRS P = .988, HR=1.00 95%CI: 1.09-0.92) and interaction effects (LRS P = .575, HR=1.03, 95% CI= 1.16-0.92). The proportional hazards assumption for this analysis was confirmed.

Missing data

We next assessed the association between missing data and depression risk group or sequential CES-D examination (Figure 3a). Logistic regression analysis based

on the data in Figure 3a indicated that depression risk group (OR=1.17; 95% CI = 1.13 to 1.21; P <.001) and sequential examination (OR=1.45; 95% CI = 1.44 to 1.46; P < .001) were both statistically significantly associated with missing CES-D data. Figures 3b-d show the proportion of participants completing the CED-D by depression risk and treatment groups. Logistic regression analysis by depression risk, controlling for sequential examination, indicates that, compared with those receiving the placebo, tamoxifen treatment was associated with higher proportions of missing data in the low-risk (OR=1.11; 95% CI = 1.06 to 1.16; P < .001) and the medium-risk groups (OR=1.12; 95% CI = 1.04 to 1.21; P < .001), but not in the high-risk group (OR=0.99; 95% CI = 0.84 to 1.16; P = .91). If tamoxifen-associated depression was the primary cause of these missing data, we would have predicted a positive (dose-response) increase in the magnitude of the odds ratios from the lowest to the highest depression risk group.

We noted in our previous paper *(15)* that it was difficult to continue to collect quality of life data after a participant had gone off treatment. However, participants in the P-1 study were asked about their primary reason for going off treatment and their responses were recorded on an Off Therapy Form (OTF) that included "depression" as one of 10 specific response categories.

Of the 11,064 participants in this cohort, we collected an OTF for 3,539 (80.7%) of 4,382 women who missed at least one CES-D examination. The presence of an OTF showed a moderate positive correlation with the total number of missing CES-D examinations (r=0.62; P < .001). The women who completed an OTF accounted for 12,693 (89.7%) of 14,149 missing CES-D examinations. Only 110 (3.1%) of these women reported that depression was the primary reason for going off therapy. The most frequent reasons for going off therapy were non-medical in nature (1667 women,

[47.1%]), perceived toxicities (921 women, [26.0%]), and various protocol and non-protocol medical conditions (841 women, [23.8%]). Table 5 shows the distribution of women who reported that depression was their primary reason for going off treatment by treatment group and depression risk group. An analysis of these data using binary logistic regression found a statistically significant effect for depression risk group (OR=2.37; 95% CI = 1.83 to 3.07; $P < .001$) and a statistically nonsignificant effect for treatment group (OR=1.10; 95% CI = 0.75 to 1.62; $P = .63$), indicating that the cases of depression that lead women to quit their assigned treatment did not occur with a greater frequency in those in the tamoxifen arm.

DISCUSSION

Tamoxifen is the most widely prescribed anti-cancer agent currently in use. It has been proven to be effective against breast cancer as an adjuvant treatment and in a preventative setting *(12, 31)*. Given the widespread use of tamoxifen, it is important to fully investigate all of the potential side-effects that may be associated with its administration, so that women, together with their physicians, can make an informed decision regarding its potential costs and benefits and its appropriateness for their individual situations.

This study is an extension of our earlier report *(16)* on the HRQL data from the NSABP P-1 study. Previously, we found no evidence for an association between tamoxifen treatment and depression in the overall P-1 study cohort. In this study, we recognized that vulnerability to clinically identifiable depressive disorders is not randomly distributed in the general female population and that the effects of tamoxifen on susceptible women in the P-1 study may have previously gone undetected.

Our initial problem was the *a priori* identification of subgroups of women with a potential clinical susceptibility for depression. Because the self-administered depression screening form (CES-D) used in the P-1 study provides information on short-term symptoms of affective distress and is not intended for use as a diagnostic instrument *(21)*, we incorporated the participants' self-reported medical history of depression, use of prescription anti-depressive medications, and experience of extended periods (greater than 12 months) of dysphoric mood to assign clinical risk. Based on these data, women were prospectively assigned to one of three depression risk groups. We hypothesized that the higher a women's depression group, the greater the likelihood that she would experience a clinically diagnosable episode of depression.

The P-1 staff were trained to check the consistency and appropriateness of the self-reported data about prior treatment for depression and the use of anti-depressant medications as a routine part of the medical screening procedure carried out during entry/eligibility interview. These procedures were designed specifically to minimize false positive classification errors. However, there was little that the interviewer could do to detect false negative classification errors in which a potential participant did not, for whatever reason, report the requested screening information. The overall effect of this inability to control for false negative classification errors for the current study was to create a potential misclassification bias in which women at increased risk for depression may have been placed, at an unknown rate, in one of the lower risk groups. Although less than ideal, the effect of this bias is conservative in nature, operating to maintain the comparative validity of the most important high-risk depression group.

We found a statistically significant dose-response relationship between the level of the depression risk group (high>medium>low) and the proportion of the women in

each depression risk group who scored above the clinical cut-off of ≥16 on the CES-D at baseline and at every follow-up interview. In addition, women in the higher risk groups (high>medium>low) scored above the clinical cut-off on a greater proportion of their follow-up interviews and, on average, had higher maximum CES-D scores. Together, these data suggest that there was a dose-response effect in which women in the higher depression risk groups (high>medium>low) were more likely to experience a clinically significant episode of affective distress and that these episodes, on average, were more persistent and severe than the episodes in the lower risk groups. Finally, we found that the distribution of social and demographic correlates (i.e., age, marital and employment status, education, use of medical services) across the three depression risk groups defined in this study followed the same general patterns of risk previously identified in the ECA study of depression among the general population *(17)*. All of the above findings serve to support the validity of the risk assignments used in our study.

The primary test of our research question involved stratifying each depression risk group by treatment assignment (tamoxifen versus placebo) and comparing the corresponding proportions of women at each follow-up interview who scored above the clinical cut-off ≥16 on the CES-D. We found no effect of tamoxifen for any of the three depression risk groups.

Besides the lack of a positive association between tamoxifen use and depression, there are at least two possible alternative explanations for our negative findings: lack of statistical power and missing data. We carried out a post-hoc effect size analysis to determine the size of the difference between the treatment arms that might have been detected. For our highest risk depression group (n=519), we had an 80% chance of detecting at least a 37% (OR ≥ 1.37) increase between the two study

arms in the proportions of women scoring above the CES-D clinical cut-off of ≥16 at any single examination point. When a repeated measures design was used, we had sufficient power to detect a mean increase of 24% (OR ≥ 1.24) in the proportion of women in either arm scoring above the CES-D clinical cut-off *(32-33)*. We considered these to be acceptable levels of statistical power for the identification of clinically significant treatment effects in our high-risk depression group. The detectable odd ratios were, of course, even smaller for the low- and medium-risk depression groups.

We also assessed the contribution of missing data to explain the negative association between tamoxifen and depression in the P-1 study. An initial analysis showed that assigned depression risk was statistically significantly associated with missing data rates over the course of the study. If a "tamoxifen-associated" depression was the primary cause of these rates, we would have predicted that the tamoxifen treatment group in the higher depression risk groups would show a progressively greater differential off-treatment rate compared with the placebo group. This expectation was not confirmed by our data for the high-risk depression group.

In addition, we also examined the reasons given for going off the assigned treatment. There was a strong statistical association in the P-1 study between stopping assigned treatment and missing HRQL data *(16)*. An analysis of the reasons for going off treatment in 81% of the women with missing HRQL data resulted in the following observations: (a) depression was cited as a relatively infrequent reason for going off treatment; (b) the higher the depression risk group, the greater the likelihood that depression was cited as the reason for going off treatment; and, (c) within each depression risk group, depression was cited as the reason for going off treatment by similar proportions of women, regardless of treatment assignment. A separate paper

(34) has implemented a sensitivity analysis on these data with equally negative results. The findings in this report together with this sensitivity analysis , indicates that there are no clear patterns in the missing data that serve to undermine the conclusions drawn from our primary analysis.

The results of our analysis strengthen our previous conclusion regarding lack of evidence for an association between tamoxifen use and depression in the P-1 data by provisionally extending our findings to subgroups of women at a high-risk for clinically identifiable episodes of depression. Clinically, these findings have two major implications. First, the evidence from NSABP's P-1 study does not lend support to the idea that that tamoxifen should be considered a causal risk factor for the onset of depressive symptoms and/or the prolongation of depressive episodes that occur among treated women. Second, the findings of this study suggest that physicians need not automatically disqualify women as candidates for tamoxifen treatment simply because they report a history of depressive symptoms or prior treatment for a depressive disorder. Nevertheless, it is still essential that physicians carefully screen for affective disorders and treat or refer potential cases of depression encountered in routine clinical practice.

Finally, there are two important limitations on these conclusions that require discussion, one statistical and the other methodological. Statistically, it was the large size of the P-1 study that permitted us to identify and carry out stratified analyses of groups of women with a differential risk for depression. However, we also noted that there were limits on our statistical power to detect an increase in the proportion of women reporting clinically significant levels of depressive symptoms on the CES-D, particularly in the high-risk depression group. For this reason, we cannot absolutely

exclude the possibility that there may be rare cases in which women react negatively to tamoxifen treatment with potentially life-threatening depressions. Here its is useful to recall that data on neuro-mood toxicities were collected for P-1 participants and periodically reviewed as part of the routine safety monitoring procedures. Over the full course of the P-1 study, there were a total of three suicides, one from the placebo and two from the tamoxifen-treated group, and there were no statistically significant differences in the distribution of women reporting suicidal ideation across the two trial arms.

The methodological limitations of this report (i.e., the lack of standardized psychiatric diagnoses and missing HRQL data) are primarily due to the fact that the goals of this study were secondary to the main clinical objectives that determined the design of the P-1 study. A more definitive analysis would require additional data from a potentially smaller, yet more focused study in which an investigation of the relationship between clinical depression and tamoxifen treatment was the primary scientific objective. Such a study would have to have the following minimum features: (a) a double-blind, placebo-controlled, randomized design; (b) participants who are at high risk for breast cancer, rather than breast cancer patients (to avoid potential confounding due to clinical diagnosis and treatment); (c) participants who are stratified on a reliable measure of risk for affective disorder (e.g., lifetime diagnosis, Schedule for Affective Disorders and Schizophrenia - Lifetime version); (d) periodic administration, in whole or in part, of a standardized psychiatric diagnostic instrument (e.g., Diagnostic Interview Schedule) by a trained interviewer; and (e) continued collection of the psychiatric interview data even if the participant goes off the assigned treatment for any reason, except death or consent withdrawal. Whether the additional information obtained from

such a study would justify the time and the expense involved in its collection is a problematic question that is beyond the scope of this report.

Acknowledgements

This investigation was supported by public health service grants from the National Cancer Institute (NCI-U10-CA-37377/69974) and a career development award from the Department of Defense to Dr. Day (DAMD17-97-1-7058). Dr. Ganz is supported in part by an American Cancer Society Clinical Research Professorship. The authors wish to thank Samuel Wieand, Ph.D, Stephanie Land, Ph.D., and Ms. Sheela Goshal of the NSABP Biostatistical Center and D. Lawrence Wickerham, M.D., of the NSABP Operations Center, for their help in the preparation of this report.

REFERENCES

1. Cathcart CK, Jones SE, Pumroy CS, Peters GN, Knox SM, Cheek JH. Clinical recognition and management of depression in node negative breast cancer patients treated with tamoxifen. Breast Cancer Res Treat 1993;27:277-81.
2. Shariff S, Cumming CE, Lees A, Handman M, Cumming DC. Mood disorder in women with early breast cancer taking tamoxifen, an estradiol receptor antagonist: An unexpected effect? Ann NY Acad Sci 1995;761: 365-8.
3. Anelli TF, Anelli A, Tran KN, Lebwhol DE, Borgen PI. Tamoxifen administration associated with a high rate treatment limiting symptoms in male breast cancer patients. Cancer 1994;74:74-7.
4. Pluss JL, Dibella NJ. Reversible central nervous system disorder due to tamoxifen in a patient with breast cancer. Ann Intern Med 1984;101:652.
5. Duffy LS, Greenberg DB, Younger J, Ferraro MG. Iatrogenic acute estrogen defficiency and psychiatric syndromes in breast cancer patients. Psychosomatics 1999;40:304-8.
6. Physicians Desk Reference (53rd ed.), Oradell, NJ: Medical Economics Co., 1999.
7. Love RR, Cameron L, Connell BL, Leventhal H. Symptoms associated with tamoxifen treatment in postmenopausal women. Arch of Int Med 1991;151:1842-7.
8. Fallowfield L, Fleissig A, Edwards R, West A, Powles TJ, Howell A, et al. Tamoxifen for the prevention of breast cancer: psychosocial impact on women participating in two randomized controlled trials. J Clin Oncol 2001;19:1885-92.
9. Halbreich U. Role of estrogen in postmenopausal depression. Neurology. 1997;48:S16-S20.

10. Gregoire AJP, Kumar R, Everitt B, Henderson AF, Studd JW. Transdermal oestrogen for the treatment of severe postnatal depression. Lancet 1996;347:930-3.

11. Murray D. Oestrogen and postnatal depression(editorial). Lancet 1996;347:918-9.

12. Fisher B, Costantino JP, Wickerham L, Kavanah, M, Cronin WM, Vogel, V et al: Tamoxifen for the prevention of breast cancer: a report from the NSABP P-1 study. J Natl Cancer Inst 1998;90:1371-88.

13. Fisher B, Costantino J. Highlights of the NSABP breast cancer prevention trial. Cancer Control 1997;4:78-86.

14. Ganz PA, Day R, Ware JE, Redmond C, Fisher B. Base-line quality-of-life assessment in the National Surgical Adjuvant Breast and Bowel Project breast cancer prevention trial. J Natl Cancer Inst 1995;87:1372-82.

15. Ganz PA, Day R, Costantino JP. Compliance with quality of life data collection in the NSABP breast cancer prevention trial. Stat Med 1998;17:613-22.

16. Day R, Ganz PA, Costantino JP, Cronin W, Wickerham DL, Fisher B. Health-Related Quality of Life and Tamoxifen in Breast Cancer Prevention: A Report from the NSABP P-1 Study. J Clin Oncol 1999;17:2659-69.

17. Weissman MM, Bruce ML, Leaf PJ, Florio LP, Holzer C. Affective disorders. In: Robins LN, Regier DA (eds.) Psychiatric Disorders in America: The Epidemiological Catchment Area Study. New York: The Free Press, 1991.

18. Diagnostic and Statistical Manual of Mental Disorders: DSM-IV. Washington, DC: American Psychiatric Association:2000.

19. Spitzer RL, Endicott J, Robbins E. Research Diagnostic Criteria. NY: Biometrics Research Division, New York State Psychiatric Institute, 1978.

20. Dohrenwend BP, Shrout PE, Egin E, Mendlesohn FS. Nonspecific psychological distress and other dimensions of psychopathology: Measures for use with the general population. Arch Gen Psychiatry 1980;37:1229-36.

21. Radloff LS. The CES-D scale: A self-report depression scale for use in the general population. Applied Psychological Measurement 1977;1:385-401.

22. Myers JK, Weissman MM. Use of a self-report symptom scale to detect depression in a community sample. Am J of Psychiatry 1980;137:1081-4.

23. Lin N, Dean A, Ensel WM. Social support, Life Events, and Depression. New York: Academic Press, 1986.

24. Boyd JH, Weissman MM, Thompson WD, Myers JK. Screening for depression in a community sample: Understanding discrepancies between depression symptom and diagnostic scales. Arch Gen Psychiatry 1982;39:1195-200.

25. Spitzer RL, Endicott J. Schedule for Affective Disorders and Schizophrenia. New York: Biometrics Research Division, New York state Psychiatric Institute, 1978.

26. Kaelber CT, Moul DE, Farmer ME. Epidemiology of depression. In; Beckham EE, Leber WR (eds.) Handbook of Depression (2nd edition). NY:Guilford Press, 1995. 3-35.

27. Keller MB, Lavori PW, Rice J, Coryell W, Hirschfeld RM. The persistent risk of chronicity in recurrent episodes of nonbipolar major depressive disorder: A prospective follow-up. Am J Psychiatry 1986;143:24-8.

28. Keller MB, Lavori PW, Collins E. Predictors of relapse in major depressive disorder. JAMA 1983;250:3299-304.

29. Leaf PJ, Livingston MM, Tischler GL, Weissman M, Holzer C, Myers JK. Contact with health professionals for the treatment of psychiatric and emotional problems. Med Care 1988;23:1322-37.

30. Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst 1999;91:1541-8.

31. Fisher B, Costantino JP, Redmond C, et al. A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor -positive tumors. N Eng J Med 1989;320:479-84.

32. Fleiss JH. Statistical Methods for Rates and Proportions (2^{nd} ed). NY: John Wiley and Sons,1981.

33. Brown H, Prescott R. Applied Mixed Models in Medicine. NY: John Wiley and Sons, 1999.

34. Land S, Wieand S, Day R, Ten Have T, Costantino JP, Lang W, et al. Methodological issues in the analysis of quality of life data in clinical trials: illustrations from the National Surgical Adjuvant Breast and Bowel Project (NSABP) Breast Cancer Prevention Trial, Statistical Design, Measurement and Analysis of Health Related Quality of Life, edited by M. Mesbah, B. Cole and M. Lee, Klewer Academic Publishers (in press).

35. Stewart AL, Ware JE. Measuring Functioning and Well-Being: The Medical Outcomes Approach. Durham, NC: Duke university press, 1992.

Figure Legends

Figure 1a-d: Proportion of P-1 participants Scoring ≥16 on the CES-D with 95% Confidence Intervals by Depression Risk Groups (Low, Medium, High) Alone (1a) and by Depression Risk Group and Treatment Assignment (Placebo/Tamoxifen) (1b-d)

Figure 2: Kaplan-Meier Plot of Time from Randomization to First CES-D ≥16 by Depression Risk Group (Low, Medium, High) and Treatment Assignment (Placebo/Tamoxifen)

Figure 3a-d: Proportion of P-1 Participants Completing the QoL Questionnaire by Depression Risk Groups (Low, Medium, High) Alone (3a) and by Depression Risk Group and Treatment Assignment (Placebo/Tamoxifen) with 95% Confidence Intervals (3b-d)

Table 1.

Distribution of Self-Reported Risk Factors for Clinical Depressive Disorders at Baseline Examination

Risk Group (Items Endorsed)	Prior History of Depression	Anti-Depressant Medications	Persistent Dysphoria	Risk Factor Pattern		0-3 Risk Factors	
				n	%	n	%
Low (0)	no	no	no	7964	71.99	7964	71.99
Medium (1)	no	no	yes	621	5.61	1628	14.71
	no	yes	no	668	6.04		
	yes	no	no	339	3.06		
Medium (2)	no	yes	yes	120	1.08	953	8.61
	yes	no	yes	202	1.83		
	yes	yes	no	631	5.70		
High (3)	yes	yes	yes	519	4.69	519	4.69

Table 2

Distribution of P-1 Participants on ECA Social and Demographic Correlates of Clinically Diagnosed Depressive Disorders by Depression Risk Score

Sociodemograpic item	Depression Risk Score* Low (%)	Medium (%)	High (%)	Odds Ratio†	95% Confidence Interval on OR
Martial status: divorced /separated	11.1	17.7	23.5	1.63	1.50-1.98
Employment status: not working	4.4	7.9	12.2	1.78	1.58-2.01
Visited doctor within last 3 months	71.0	76.4	84.4	1.39	1.28-1.51
Hospitalized within last 5 years	42.7	48.6	54.9	1.27	1.19-1.36
Age: ≥ 60	29.9	27.4	24.1	0.87	0.81-0.94
Education: > High School	66.6	66.7	70.0	1.04	0.97-1.12
Income: > Median	46.1	37.6	31.5	0.72	0.67-0.77

* Depression risk groups were assigned on the basis of the participants' response to three medical history questions; prior history of depression, use of anti-depressant medication, persistent mood disturbance (dysphoria). Each positive answer was worth one point. Participants' with a score of 0 were assigned to the "low-risk" group, those with a score of 1-2 to the medium-risk group, and those with a score of 3 to the high-risk group.

† Odds ratios were determined by binary logistic regression, $P < .001$ for all groups compared with referent groups, except for education where $P = .235$

[

Table 3

Distribution of Center for Epidemiological Studies-Depression (CES-D) Scale Variables for P-1 Participants who Scored Above the Clinical Cut-off of ≥16 by Depression Risk Group

CES-D Variable	Depression Risk Group* Low	Medium	High
% follow-up exams ≥16†			
mean:	0.087	0.192	0.357
95% CI for mean:	0.083-0.091	0.181-0.203	0.325-0.389
Maximum score ≥16†			
median:	22	24	27
mean:	23.97	25.61	28.58
95% CI for mean:	23.66-24.28	25.16-26.06	27.62-29.54
All scores ≥16†			
median:	20	21	22
mean:	21.52	22.49	23.74
95% CI for mean:	21.30-21.74	22.17-22.81	23.10-24.38

* Depression risk groups were assigned on the basis of the participants' response to three medical history questions; prior history of depression, use of anti-depressant medication, persistent mood disturbance (dysphoria). Each positive answer was worth one point. Participants' with a score of 0 were assigned to the "low-risk" group, those with a score of 1-2 to the medium-risk group, and those with a score of 3 to the high-risk group.

† For all between group comparisons using Kruskal-Wallis and ANOVA: $P < .001$

Table 4

Comparison (binary logistic regression) of the Proportion of P-1 Participants in Each Treatment Group (tamoxifen versus placebo) who Scored ≥16 on the Center for Epidemiological Studies-Depression (CES-D) Scale by Depression Risk Group and Sequential Examination

		Sequential Examination						
Depression Risk Group*	Baseline	3 months	6 months	12 months	18 months	24 months	30 months	36 month
Low								
OR†	1.22	1.04	1.01	1.02	0.88	0.96	0.93	0.86
95% CI	0.96 to 1.55	0.86 to 1.25	0.85 to 1.19	0.86 to 1.02	0.75 to 1.04	0.80 to 1.13	0.78 to 1.12	0.71 to 1
P	0.10	0.68	0.91	0.86	0.14	0.60	0.44	0.11
Medium								
OR†	1.03	1.29	1.10	0.99	0.91	1.04	0.96	1.01
95% CI	0.81 to 1.30	1.04 to 1.60	0.89 to 1.35	0.81 to 1.22	0.74 to 1.13	0.82 to 1.30	0.75 to 1.22	0.79 to 1
P	0.84	0.02	0.39	0.95	0.40	0.76	0.72	0.94
High								
OR†	0.89	0.78	0.74	0.62	0.84	0.83	1.00	1.00
95% CI	0.61 to 1.30	0.54 to 1.14	0.50 to 1.09	0.41 to 0.92	0.56 to 1.26	0.54 to 1.28	0.65 to 1.54	0.64 to 1
P	0.54	0.21	0.13	0.02	0.40	0.40	0.99	0.99

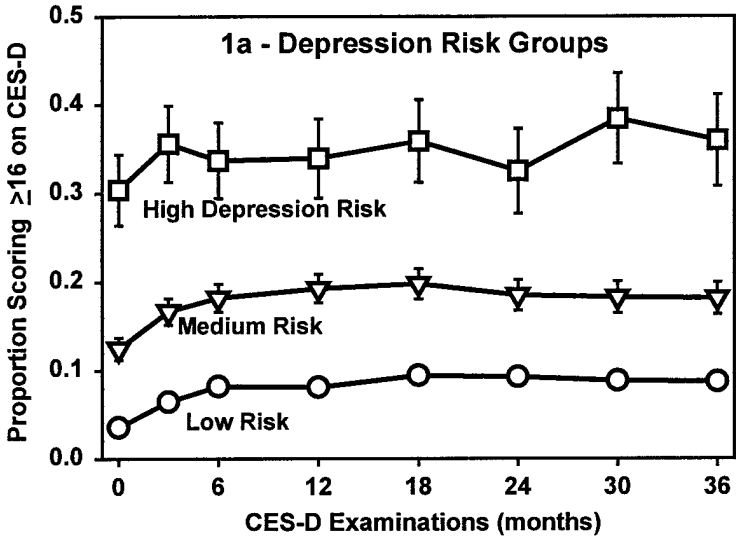
* Depression risk groups were assigned on the basis of the participants' response to three medical history questions; prior history of depression, use of anti-depressant medication, persistent mood disturbance (dysphoria). Each positive answer was worth one point. Participants' with a score of 0 were assigned to the "low-risk" group, those with a score of 1-2 to the medium-risk group, and those with a score of 3 to the high-risk group.

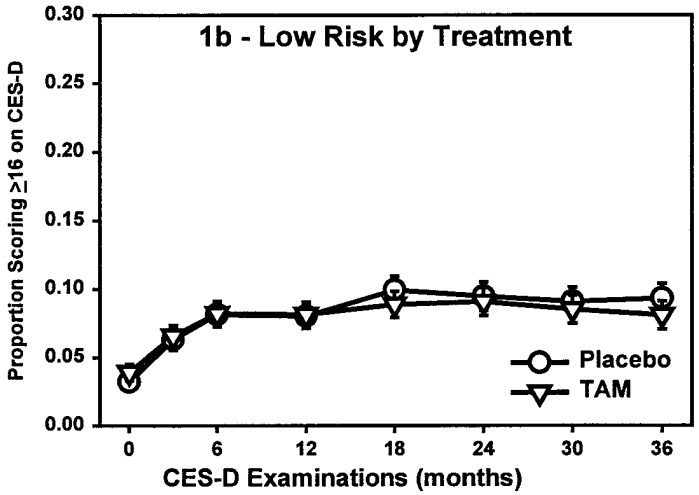
† OR = odds ratio (OR >1.0 indicates a greater proportion of women in the tamoxifen group); CI = confidence interval

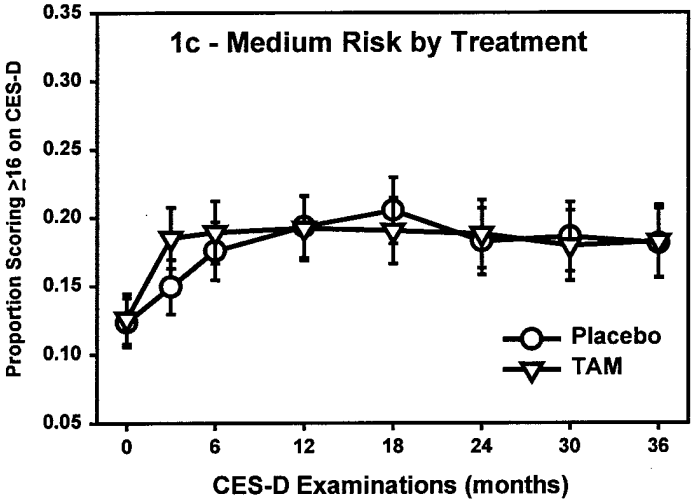
Table 5
Reasons Cited for Going Off Treatment
by Depression Risk* and Treatment Group

Reasons cited for going off treatment	Low risk		Medium risk		High risk		Overall
	placebo	tamoxifen	placebo	tamoxifen	Placebo	tamoxifen	
Depression (no. participants)	20	27	21	24	9	9	110
other reasons (no. participants)	1130	1275	416	431	83	94	3429
depression as % of all off treatment reasons	1.7	2.1	4.8	5.3	9.8	8.7	3.1

* Depression risk groups were assigned on the basis of the participants' response to three medical history questions; prior history of depression, use of anti-depressant medication, persistent mood disturbance (dysphoria). Each positive answer was worth one point. Participants' with a score of 0 were assigned to the "low-risk" group, those with a score of 1-2 to the medium-risk group, and those with a score of 3 to the high-risk group.


1a - Depression Risk Groups
Proportion Scoring ≥16 on CES-D
0.5
0.4
0.3
0.2
0.1
0.0
High Depression Risk
Medium Risk
Low Risk
0
6
12
18
24
30
36
CES-D Examinations (months)


1b - Low Risk by Treatment
Proportion Scoring ≥16 on CES-D
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Placebo
TAM
0
6
12
18
24
30
36
CES-D Examinations (months)


1c - Medium Risk by Treatment
Proportion Scoring ≥16 on CES-D
0.35
0.30
0.25
0.20
0.15
0.10
0.05
Placebo
TAM
0
6
12
18
24
30
36
CES-D Examinations (months)

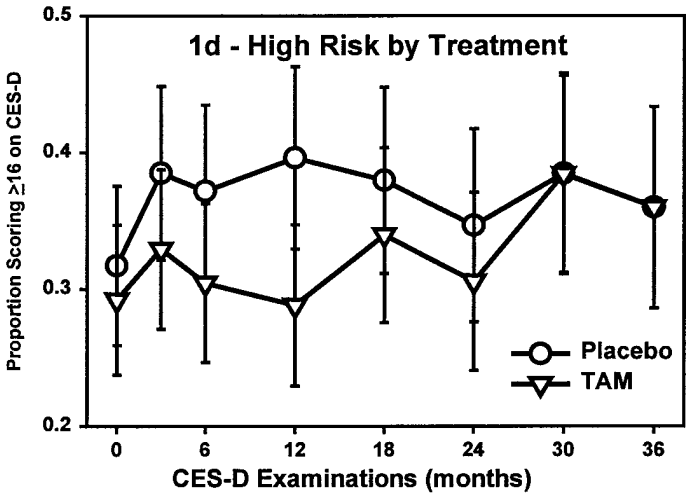
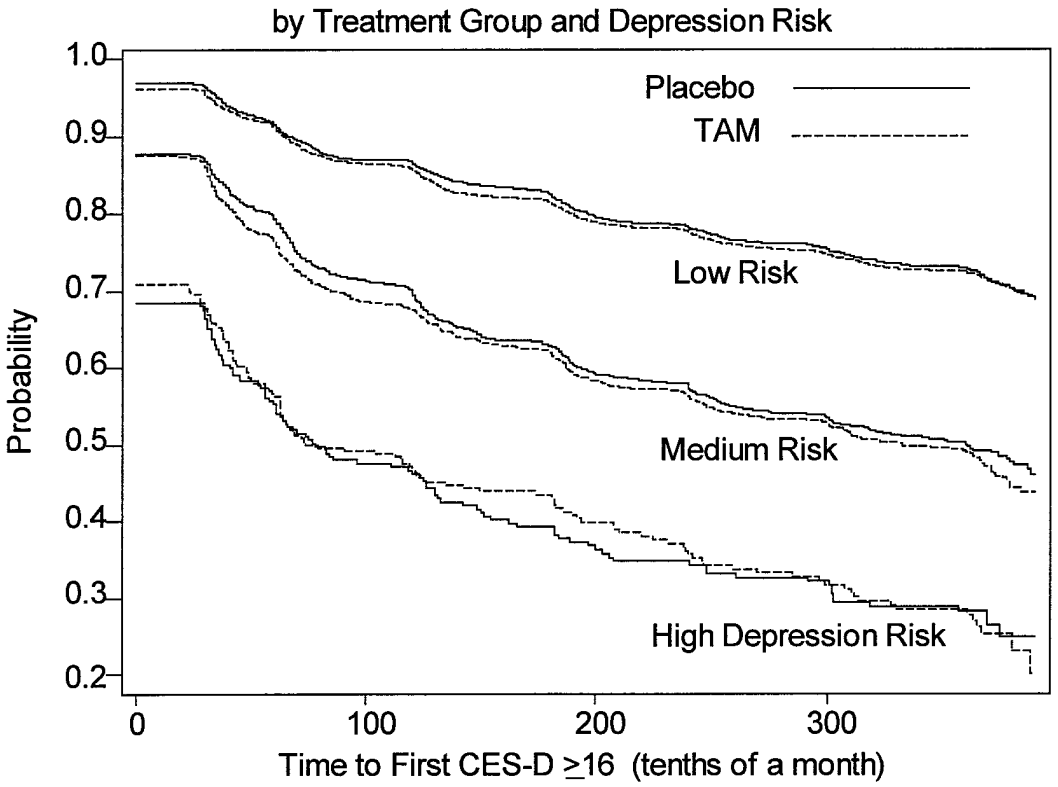
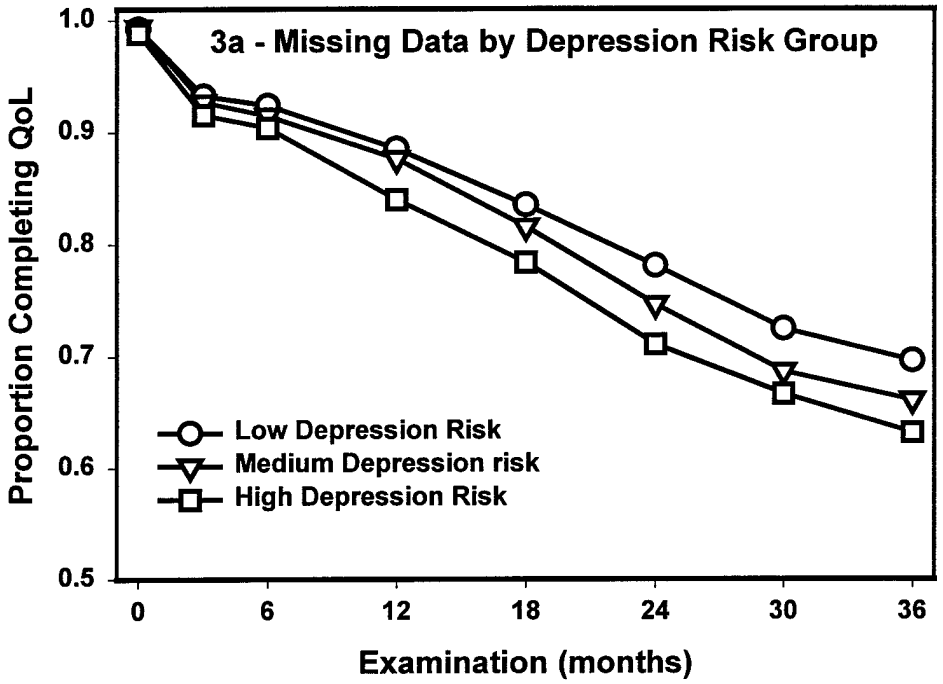
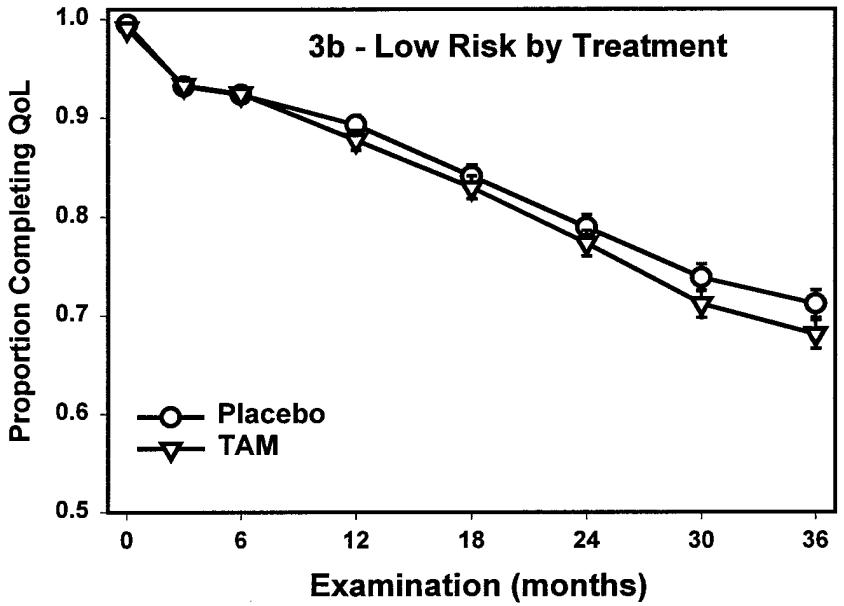

1d - High Risk by Treatment
Proportion Scoring ≥16 on CES-D
0.5
0.4
0.3
0.2
Placebo
TAM
0
6
12
18
24
30
36
CES-D Examinations (months)

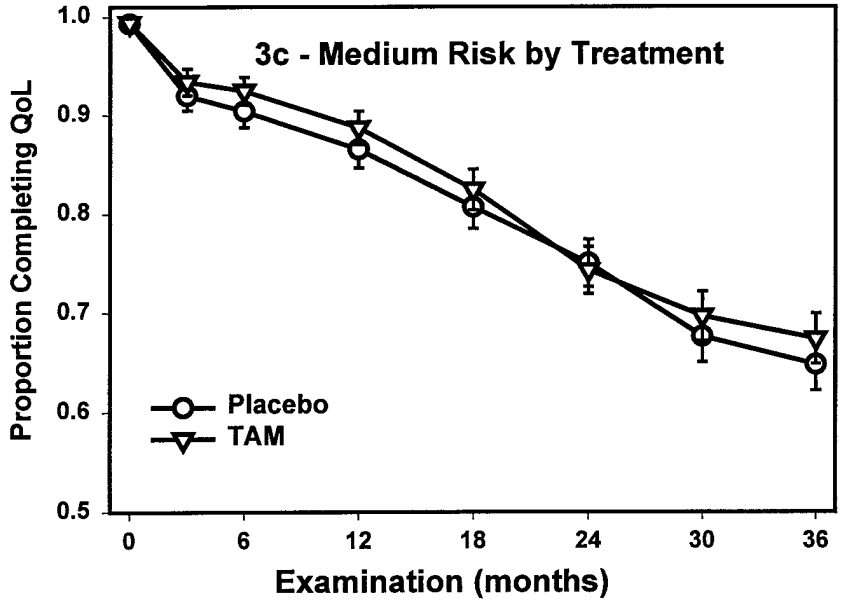
Figure 3

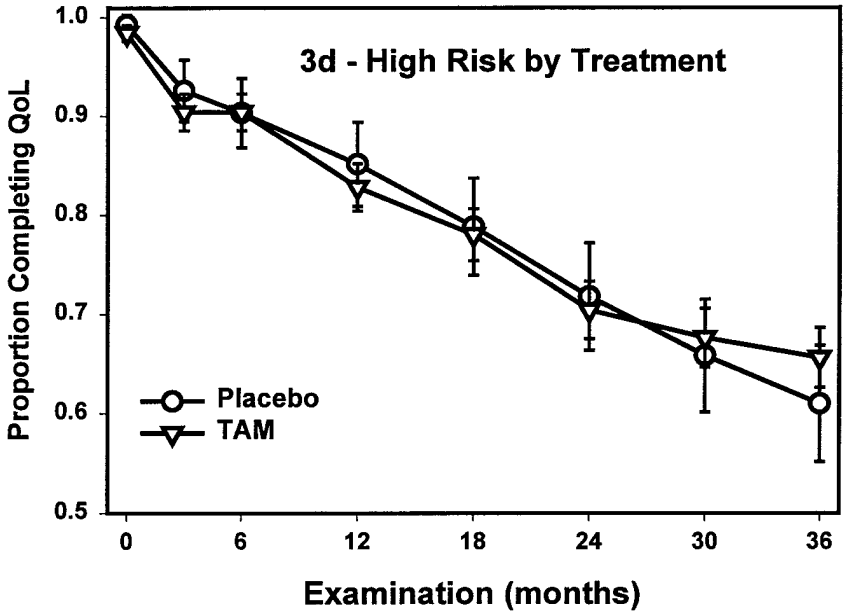


10 months:
Low Risk: TAM Sur %=.864 95%CI: .853-.875 at risk: 3159 (Plac Sur%: .870 at risk: 3190)
Med Risk: TAM Sur.%=.685 95%CI: .659-.711 at risk: 799 (Plac Sur%: .713 at risk: 863)
High Risk: TAM Sur.%=.488 95%CI: .427-.549 at risk: 123 (Plac Sur%: .475 at risk: 108)
30 months:
Low Risk: TAM Sur %=.746 95%CI: ..732-.760 at risk: 2233 (Plac Sur%: .753 at risk: 2326)
Med Risk: TAM Sur.%=.528 95%CI: .499-.557 at risk: 496 (Plac Sur%: .535 at risk: 544)
High Risk: TAM Sur.%=.317 95%CI: .258-.376 at risk: 61 (Plac Sur%: .316 at risk: 59)


3a - Missing Data by Depression Risk Group
Proportion Completing QoL
1.0
0.9
0.8
0.7
0.6
0.5
0
6
12
18
24
30
36
Examination (months)
Low Depression Risk
Medium Depression risk
High Depression Risk


3b - Low Risk by Treatment
Proportion Completing QoL
1.0
0.9
0.8
0.7
0.6
0.5
0
6
12
18
24
30
36
Examination (months)
Placebo
TAM


3c - Medium Risk by Treatment
Proportion Completing QoL
1.0
0.9
0.8
0.7
0.6
0.5
0
6
12
18
24
30
36
Examination (months)
Placebo
TAM


3d - High Risk by Treatment
Proportion Completing QoL
1.0
0.9
0.8
0.7
0.6
0.5
0
6
12
18
24
30
36
Examination (months)
Placebo
TAM

Appendix 2

Day R, Quality of life and tamoxifen in breast cancer: a summary of the findings from the NSABP P-1 study. Annals of the New York Academy of Sciences (2001, in press)

Abbreviated Title: Quality of Life in a Breast Cancer Prevention Trial

Quality of Life and Tamoxifen in a Breast Cancer Prevention Trial: A Summary of Findings From the NSABP P-1 Study

Richard Day, Ph.D.

Department of Biostatistics,

Graduate School of Public Health

University of Pittsburgh, Pittsburgh

(412) 624-4077 (phone) (412) 624-9969 (fax)

DAY@NSABP.PITT.EDU

Key Words: Quality of Life, Tamoxifen, breast cancer, prevention

Abstract

This report contains a brief summary of the health-related quality of life findings for 11,064 women taking part in the National Surgical Adjuvant Breast and Bowel Project's P-1 trial. Women taking part in this trial of tamoxifen versus placebo for breast cancer prevention were $\geq$ 35 years old and predominately white, well-educated, and middle-class, with a strong professional and technical orientation. Key findings included a lack of difference between the tamoxifen and placebo arms with regard to depression, overall physical or mental quality of life, or weight gain. The tamoxifen arm did show consistent increases in vasomotor (hot flashes) and gynecological (vaginal discharge) symptoms, as well as difficulties in certain domains of sexual functioning. It is concluded that an informed discussion with a woman considering tamoxifen therapy should include these points in the risk-benefit discussion.

Introduction

This is a brief summary of the findings from the health-related quality of life (HRQL) component of the National Surgical Adjuvant Breast and Bowel Project's (NSABP) P-1 trial, a mult-center, double-blinded, placebo controlled clinical trial designed to evaluate whether 5 years of tamoxifen therapy would reduce the incidence of invasive breast cancer in women at an increased risk for the disease. Detailed descriptions of the rationale, planning and design of the of the P-1 study and its HRQL component, as well as specific instruments, are available in separate reports.[1-5]

Subjects and Instruments

This summary focuses on the baseline HRQL examination and the first 36 months of follow-up data on 11,064 women recruited over the first 24 months of the study. The P-1 HRQL Questionnaire was composed of the Center for Epidemiological Studies – Depression Scale (CES-D), the Medical Outcomes Study (MOS) Short Form (SF-36), the MOS sexual functioning scale, and a symptom checklist (SCL). The questionnaire was to be administered to all participants prior to randomization (baseline), at 3 months and at each succeeding 6 month examination.

Results

The participants in the P-1 study were predominately white (96%); well-educated (65% $\geq$ some college), married (70%), professional and technically trained (68.2%) women, who were currently employed (64.9%) and reported a middle to upper middle class family income (median $35-49,999).

Figure 1 shows the overall proportion and total numbers of women completing the HRQL questionnaire at each examination. It provides a measure of comparative participant adherence with regard to the HRQL questionnaire in the two trial groups.

Analysis of sociodemographic and medical variables indicated that participants failing to complete the HRQL questionnaire in each group were similar cohorts of women.

Figure 2 shows the proportion of P-1 participants, by group and examination, scoring above the most frequently used clinical cut-off ($\geq$16) on the CES-D.[6-7] The youngest age group (35-49 yrs.) in both trial groups consistently had the highest proportion of members scoring above the clinical cut-off, followed by the 50-59 yrs. age group . Similar findings with regard to the relationship between the two trial groups emerged from the analysis of the 5-item mental health subscale on the MOS SF-36 (not shown).

The SF-36 results are summarized in Figure 3 using the physical and mental component scores (PCS, MCS).[8] Mean PCS declines across the age groups. On follow-up examinations, the tamoxifen group was consistently lower on the PCS only in the 50-59 yrs. age group (one-sided sign test, p=0.065); however, the absolute differences were very small, approximating 1/10 of a standard deviation. No consistent differences emerged on the MCS between the two trial groups.

Table 1 provides information on the proportion of women in the tamoxifen and placebo groups reporting symptoms on the SCL at least once during the period that the participants were on treatment – i.e., the period excluding baseline, but including the 7 follow-up examinations. The 5 symptoms with the greatest relative difference between the two trial groups are given for each age group and the 10 symptoms with the greatest relative difference are presented for all participants combined.

Figure 4 summarizes the information from the 5 items on the MOS sexual functioning scale. Plate A on Figure 4 shows that a greater proportion of participants in the tamoxifen as compared to the placebo group reported being sexually active during

the 6 months prior to each follow-up examination. Although apparently consistent, the absolute difference was small (mean=0.78%) and the findings may have been due to chance. Plates B-E show that a small, but consistently larger percentage of participants in the Tamoxifen group reported a definite or serious problem in three of the four specific domains of sexual functioning during the follow-up period.

Discussion

The cohort of women taking part in the P-1 study were not representative of the general population. They were predominately white, well-educated, and middle-class, with a strong professional and technical orientation. The initial HRQL findings presented in this report must be assessed within the context of the socioeconomic and cultural characteristics of the P-1 study cohort.

Although 31.5% our participants were did not complete the 36 month HRQL follow-up examination, we have shown that there is only a small difference in the proportion of non-adherent participants in the tamoxifen and placebo groups and that the non-adherent women in both trial groups are generally similar on key demographic, clinical and HRQL variables. Given these considerations, it seems unlikely that that a maximum difference of 3% in the HRQL follow-up rates between the two groups was sufficient to create a significant bias in our between group comparisons.

Concern has been expressed regarding the possible relationship between tamoxifen use and the onset of depression.[9-13] Women reporting a history of depressive episodes or a history of treatment for nervous or mental disorders were not excluded from the trial. If tamoxifen use was associated with the onset of clinically diagnosable depression, we would have expected to see a consistent excess of individuals scoring $\geq$16 on the CES-D in the tamoxifen group. No such consistent excess was observed.

The MOS SF-36 served in this study as a measure of overall health-related quality of life. We presented data from this instrument in terms of two high-level component scores (PCS and MCS), neither one of which demonstrated any clinically significant differences between the tamoxifen and placebo groups.

The first signs of consistent differences between the tamoxifen and placebo groups were observed in the symptom checklist (SCL). The differences between the trial groups tended to be associated with the types of vasomotor, gynecological, and sexual functioning symptoms previously reported for tamoxifen.[10, 14-15]

The data from the MOS sexual functioning scale indicate that relatively small (<4.0%), but consistent differences exist between the two groups with regard to the proportion of women reporting definite or serious problems in at least three specific domains of sexual functioning – sexual interest, arousal, and orgasm. These problems do not appear to be age group specific. Despite these findings for specific domains of functioning, there is no evidence that these problems result in a reduction in the overall proportion of women in the tamoxifen group who are sexually active.

Based on these data, we would conclude that tamoxifen use is associated with an increase in specific vasomotor, gynecological, and sexual functioning symptoms. At the same time, we did not observe any evidence that overall physical or emotional well being were significantly affected by these differences in the frequency of symptoms. We also found no evidence on the CES-D or the SF-36 mental health scale for an association in any age group between tamoxifen use and an increase in the proportion of women reporting clinically significant levels of depression.

How should clinicians integrate these research results into decision-making and recommendations to women considering the use of tamoxifen in the setting of

prevention? Many symptoms experienced by women who participated in this study are age and menopause related, and exist independent of the use of tamoxifen. However, several symptoms are substantially more frequent in women using tamoxifen and these include vasomotor symptoms (cold sweats, night sweats, hot flashes), vaginal discharge, and genital itching. Women need to be informed of these possible symptoms. Weight gain and depression, two clinical problems anecdotally associated with tamoxifen treatment in women with breast cancer, were not increased in frequency in this large placebo-controlled trial in healthy women. This is good news that must also be communicated to women.

An informed discussion with a woman considering tamoxifen therapy should include these points in the risk-benefit discussion. Disclosure of likely and unlikely symptoms should prepare a woman for what she might experience, and reduce her anxiety or concerns should she embark on preventive therapy. Should a woman experience untoward symptoms after starting tamoxifen treatment, the medication can be discontinued if the symptoms cannot be controlled or her personal assessment of the risks and benefits changes.

References

1. Fisher, B., J.P. Costantino, L. Wickerham, *et al.* 1998. Tamoxifen for the prevention of breast cancer: a report from the NSABP P-1 study. J. Natl. Cancer Inst. **90:** 1371-1388.

2. Fisher, B. and J. P. Costantino. 1997. Highlights of the NSABP breast cancer prevention trial. Cancer Control **4:** 78-86.

3. Ganz, P.A., R. Day, J.E. Ware, *et al.* 1995. Base-line quality-of-life assessment in the National Surgical Adjuvant Breast and Bowel Project breast cancer prevention trial. J. Natl. Cancer Inst. **87:** 1372-1382.

4. Ganz, P.A., R. Day, J.P. Costantino. 1998. Compliance with quality of life data collection in the NSABP breast cancer prevention trial. Stat Med **17:** 613-622.

5. Day, R., P.A. Ganz, J.P. Costantino, *et al.* 1999. Health-related quality of life and Tamoxifen in breast cancer prevention: A rport from the NSABP P-1 study. J. Clin. Oncol. **17:** 2659-2669.

6. Radloff, L.S. 1977. The CES-D scale: a self-report depression scale for research in the general population. Appl. Psychol. Meas. **1:** 385-401.

7. Roberts, R.E. & S. W. Vernon. 1983. The Center for Epidemiologic Studies Depression Scale: its use in a community sample. Am. J. Psychiatry **140:** 41-46.

8. Ware, J.E., M. Kosinski, S.D. Keller. 1994. SF-36 physical and mental summary scales: a user's manual: 3rd Printing Revised. Boston: The Health Institute, New England Medical Center.

9. Cathacart, C. K., S.E. Jones, C.S. Pumroy, *et al.* 1993. Clinical recognition and management of depression in node negative breast cancer patients treated with Tamoxifen. Breast Cancer Res Treat **27:** 277-281.

10. Love, R. L., L. Cameron, B.L. Connell, *et al.* 1991. Symptoms associated with Tamoxifen treatment in postmenopausal women. Arch. of Inter. Med. **151:** 1842-1847.

11. Shariff, S., C.E. Cumming, A. Lees, *et al.* 1995. Mood disorder in women with early breast cancer taking Tamoxifen, an estradiol receptor antagonist: an unexpected effect? Ann N Y Acad Sci **761:** 365-368.

12. Moredo Anelli, T., A. Anelli, K.N. Tran, *et al.* 1994. Tamoxifen administration is associated with a high rate treatment-limiting symptoms in male breast cancer patients. Cancer **74:** 74-77.

13. Pluss, J. L. & N. J. Dibella. 1984. Reversible central nervous system dysfunction due to Tamoxifen in a patient with breast cancer. Ann of Intern Med **101:** 652.

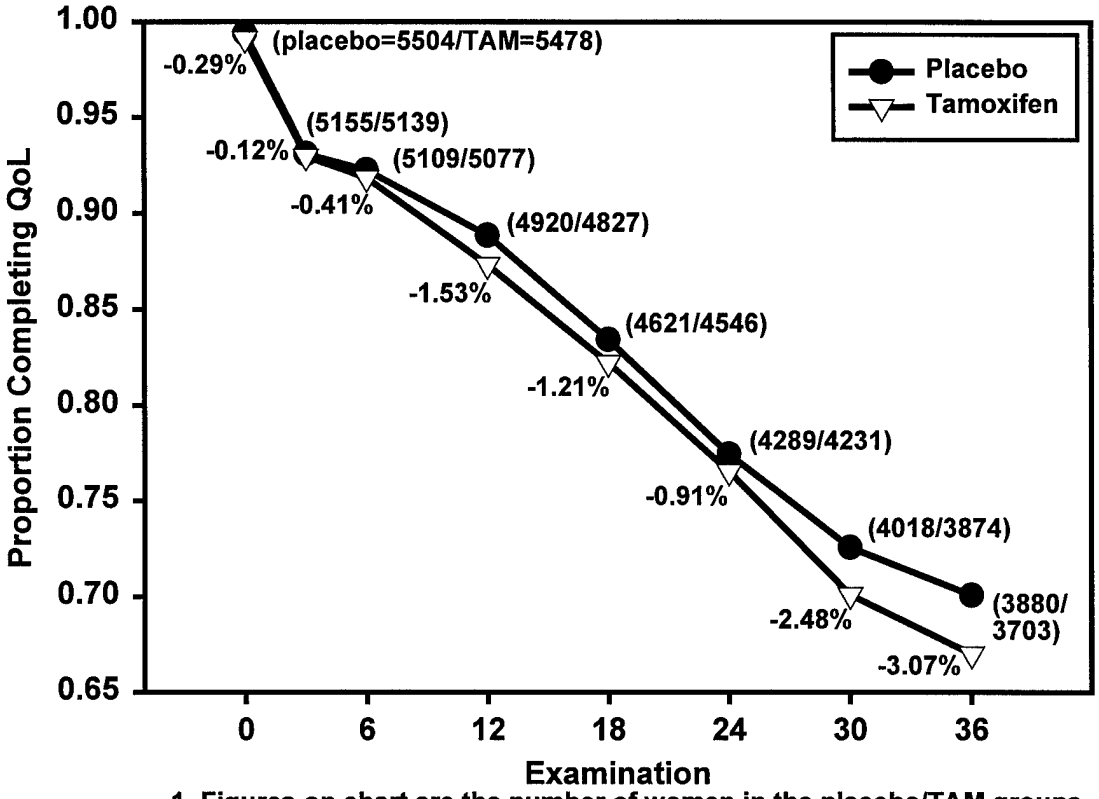
14. Fisher, B., J. Dignam, J. Bryant, et *al.* 1996. Five versus more than five years of Tamoxifen therapy for breast cancer patients with negative lymph nodes and estrogen receptor-positive tumors. J. Natl. Cancer Inst. **88:** 1529-1542.

15. Fisher, B., J. P. Costantino, C. Redmond, *et al.* 1989. A randomized clinical trial evaluating Tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors. N. Engl. J. Med. **320:** 479-484.

Table 1
Symptoms Reported At Least Once Between Months 3 and 36
With The Largest Relative Difference Between Trial Arms

Age group and Symptom	Placebo Arm Proportion (%)	Tamoxifen Arm Proportion(%)	Relative Risk (TAM/Placebo)
35-49 yrs			
1. Cold sweats	15.90	22.90	1.44
2. Vaginal discharge	46.29	62.55	1.35
3. Pain in intercourse	23.88	31.57	1.32
4. Night sweats	59.58	74.16	1.24
5. Hot flashes	65.54	81.28	1.24
50-59 yrs			
1. Cold sweats	16.11	27.00	1.68
2. Vaginal discharge	32.51	53.47	1.64
3. Genital itching	36.93	45.24	1.23
4. Night sweats	62.77	75.88	1.21
5. Bladder control (laugh)	47.67	56.94	1.19
≥60 yrs			
1. Vaginal bleeding	4.64	10.92	2.35
2. Vaginal discharge	19.82	45.81	2.31
3. Genital itching	32.05	40.96	1.28
4. Hot flashes	51.51	63.59	1.23
5. Bladder control (laugh)	49.88	56.49	1.13
Overall			
1. Vaginal discharge	34.13	54.77	1.60
2. Cold sweats	14.77	21.40	1.45
3. Genital itching	38.29	47.13	1.23
4. Night sweats	54.92	66.80	1.22
5. Hot flashes	65.04	77.66	1.19
6. Pain in intercourse	24.13	28.19	1.17
7. Bladder control (laugh)	46.65	52.51	1.13
8. Bladder control (other)	47.79	52.83	1.11
9. Weight loss	41.97	44.94	1.07
10. Vaginal bleeding	21.26	21.96	1.03

Figure 1
Proportion of Participants in the Tamoxifen and Placebo Groups Completing QoL Questionnaire by Examination[1]
(N placebo = 5537/ TAM = 5527)



1. Figures on chart are the number of women in the placebo/TAM groups completing the QoL questionnaire and the difference between TAM and placebo groups in terms of percent missing QoL data.

Figure 2
Proportion of P-1 Participants with CES-D Scores at the Level of a Potential Case (≥16) by Arm and Examination

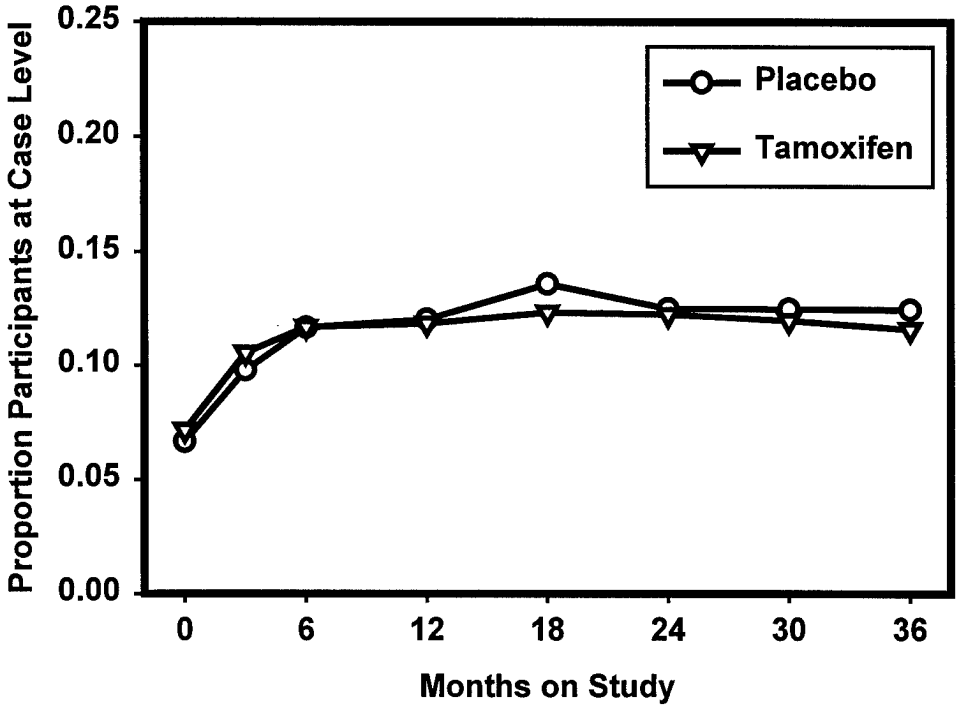


Figure 3
Mean Scores by Age Group and Examination on Sf-36 Physical and Mental Component Scores (higher scores represent better quality of life)

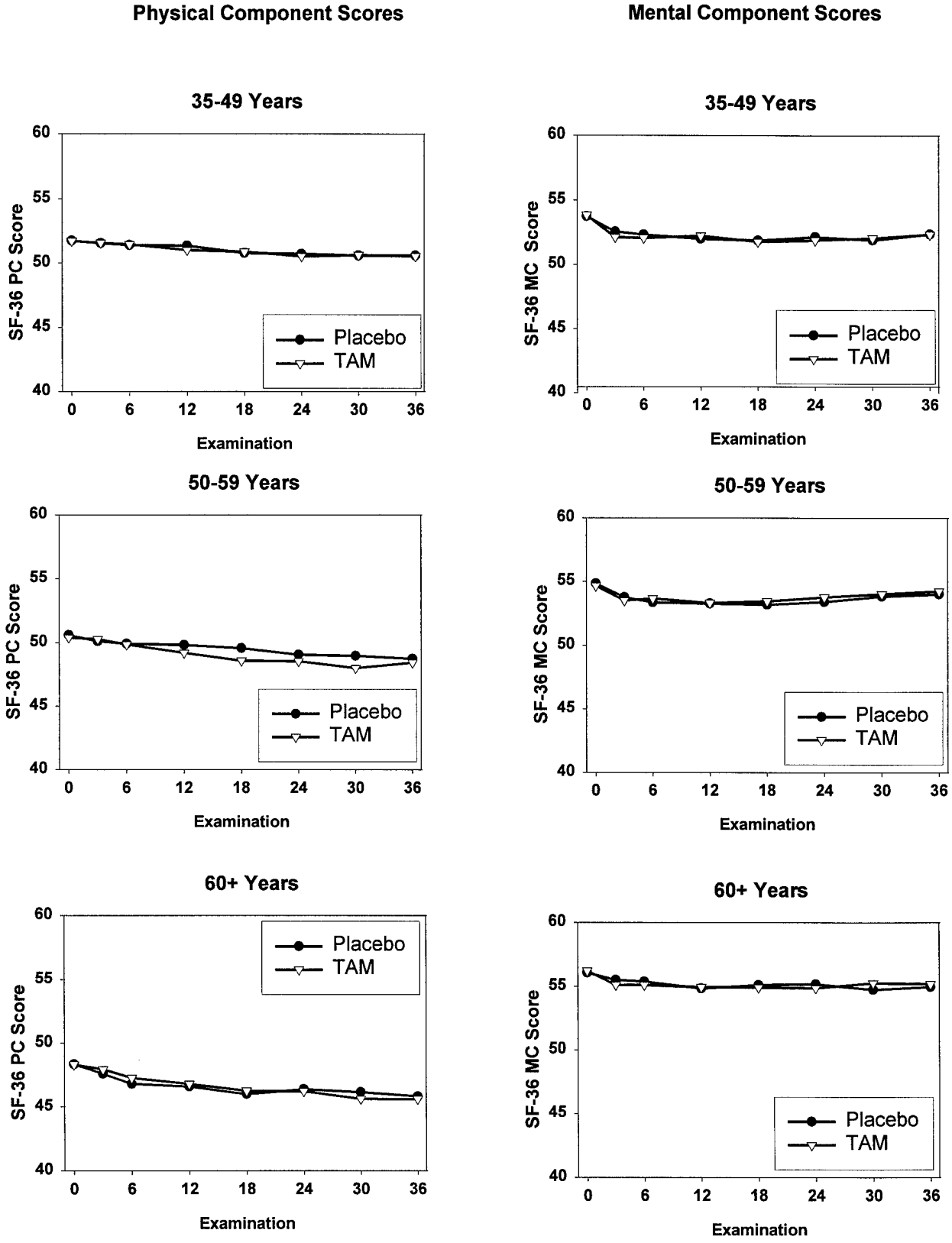
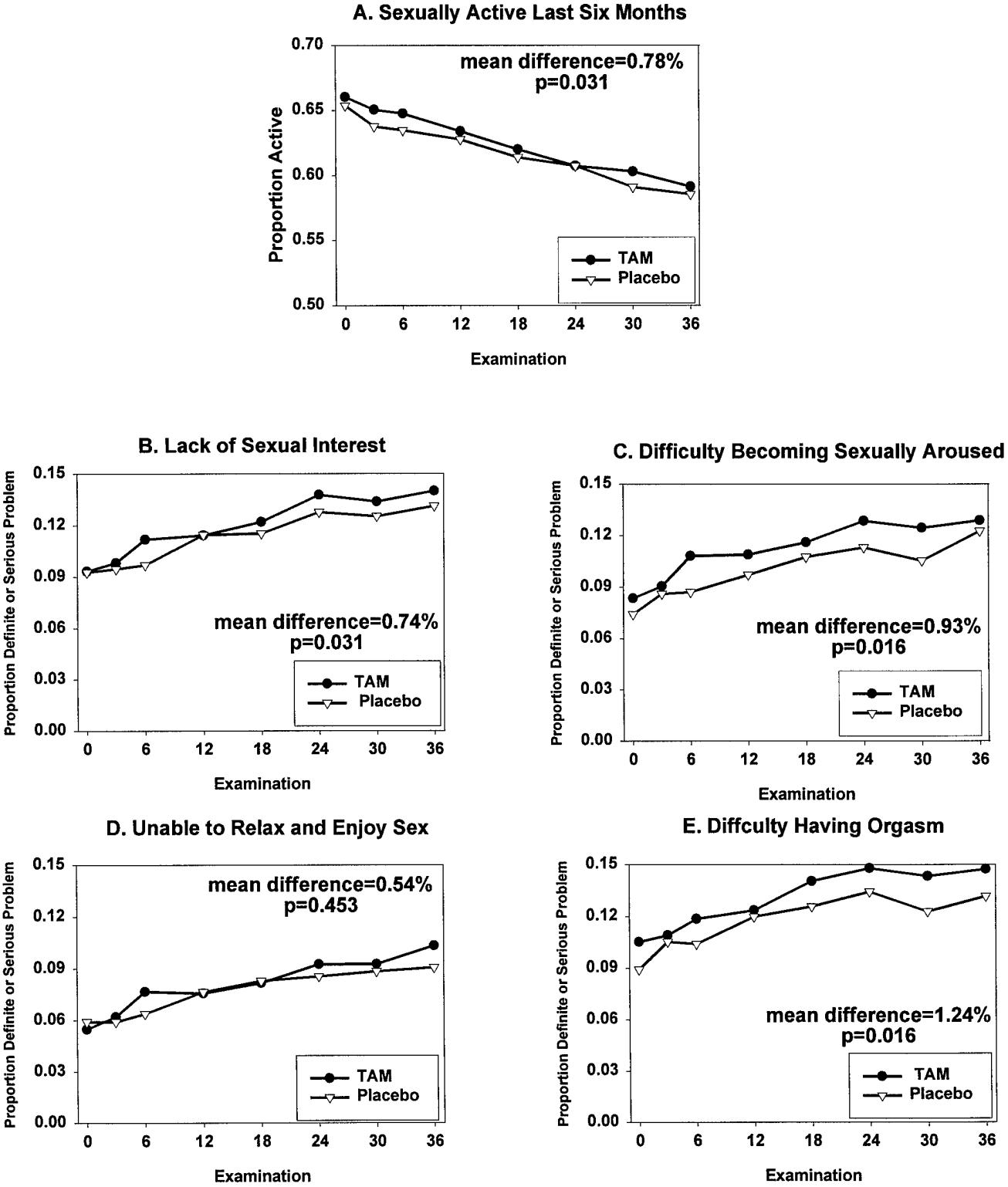


Figure 4
Proportion of Women in the Tamoxifen and Placebo Arms Reporting a Definite or Serious Problem in Past 4 Weeks on MOS Sexual Functioning Scale

(Plates B.-E. refer only to women who reported being sexually active in last 6 months)



Appendix 3

Land S, Wieand S, Day R, Have T, Costantino J, Lang W, Ganz P. Methodological issues in the analysis of quality of life data in clinical trials: illustrations from the NSABP Breast Cancer Prevention Program. In: M. Mesbah, B. Cole, M Lee (eds.), Statistical Design, Measurement and Analysis of Health Related Quality of Life. Klewer Academic Publishers (in press).

METHODOLOGICAL ISSUES IN THE ANALYSIS OF QUALITY OF LIFE DATA IN CLINICAL TRIALS: ILLUSTRATIONS FROM THE NATIONAL SURGICAL ADJUVANT BREAST AND BOWEL PROJECT (NSABP) BREAST CANCER PREVENTION TRIAL

STEPHANIE LAND[1], SAMUEL WIEAND[1], RICHARD DAY[1], TOM TEN HAVE[2], JOSEPH P. COSTANTINO[1], WEI LANG[3], AND PATRICIA A. GANZ[4]

[1] University of Pittsburgh and the National Surgical Breast and Bowel Project, [2] University of Pennsylvania School of Medicine, [3] Wake Forest University School of Medicine, [4] UCLA Schools of Medicine and Public Health and the Jonsson Comprehensive Cancer Center

We present two Quality of Life (QOL) endpoints collected in conjunction with the recently completed Breast Cancer Prevention Trial (BCPT) performed by the National Surgical Adjuvant Breast and Bowel Project. The analyses of these endpoints (depression and hot flashes) indicate the importance of randomization and give some insight about the impact of missing data in a large randomized trial.

1. Introduction

Quality of life (QOL) assessments have been increasingly included as secondary or primary endpoints in clinical trials (Tannock *et al.,* 1996; Moinpour *et al.,* 1998). The impetus for doing so comes from a desire to obtain patient-rated evaluations of treatments, especially in circumstances in which treatments have substantially differing toxicities or in which survival outcomes are not expected to be different (Ganz, 1994 a and b). Under such circumstances, an evaluation of the morbidity of treatment from the patient's or participant's perspective may in fact be the most important endpoint. Although there is now a wide range of psychometrically validated scales for the measurement of QOL in clinical trials (Cella and Bonomi, 1995), there are considerable challenges to the implementation and collection of QOL data in these studies (Bernhard *et al.,* 1998a), as well as equally formidable statistical and analytical concerns (Bernhard and Gelber, 1998b). In this paper, we provide examples from the recently completed NSABP Breast Cancer Prevention Trial (BCPT) to highlight challenges that can arise in the analyses of such data, specifically focusing on the importance of randomization and the issue of missing data and its potential to affect the interpretation of QOL outcomes.

2. Background

The BCPT was a double-blinded, placebo-controlled clinical trial that was open for accrual from June 1, 1992, through September 30, 1997. During this interval 13,338 women at high risk for breast cancer were randomly assigned to receive either 20 mg/day of tamoxifen or placebo for a duration of five years. The primary objective of the trial was to determine if tamoxifen therapy would reduce the risk of breast cancer among women. Secondary objectives related to the full benefit/risk profile of tamoxifen use in healthy women. Participants in the trial were screened for breast cancer at six-month intervals by clinical breast examination and at yearly intervals by bilateral mammography. At each screening, visit participants were also evaluated for several other endpoints including heart disease, fractures, thromboembolic disease, and endometrial cancer. Heart disease and fractures were included because it was theorized that tamoxifen might also reduce the risk of these problems. Thromboembolic disease and endometrial cancer were included

because these were known side effects associated with tamoxifen therapy. As an additional means to monitor the safety of treatment in the trial, the incidence of all invasive cancers and the occurrence of all deaths were also included as endpoints.

The results of the BCPT have been reported (Fisher *et al.,* 1998), as has a study of the risk-benefit ratio for tamoxifen (Gail *et al.,* 1999). During follow-up, 175 participants randomized to receive placebo developed invasive breast cancer compared to only 89 randomized to receive tamoxifen, indicating an estimated 50% reduction in the risk of breast cancer from the use of tamoxifen. Other major findings include the detection of a preventive effect on osteoporotic fractures, no effect on heart disease, and a confirmation of the known side effects of endometrial cancer and thromboembolic disease. These overall findings will not be discussed in this manuscript, as we wish to focus on issues that relate to QOL studies.

Because the participants in this trial were healthy women, the monitoring of their QOL during the intervention was of particular importance. Thus, the NSABP included a concurrent QOL study designed to describe side effects of tamoxifen, to examine the relationship between side effects and QOL, to compare the side effects and QOL in placebo and treated subjects, and to examine the effects of symptoms on compliance with study medication. The BCPT QOL questionnaire was a 104-item battery that included four instruments: the Center for Epidemiologic Studies Depression Scale (CES-D) (Radloff, 1997); the Medical Outcomes Study 36-Item Short Form (SF-36) (Ware *et al.,* 1994); a symptom checklist based on the Postmenopausal Estrogen Progesterone Intervention (Shumaker S., personal communication) specifically adapted for the BCPT trial; and the Medical Outcomes Study Sexual Problems Scale (Sherbourne, 1992). These instruments were selected because of their psychometric characteristics and validity, the availability of normative data in healthy women, and ease of self-administration. The latter was particularly important because the trial was conducted at several hundred clinical centers throughout North America and the battery of questions we asked was completed on multiple occasions in conjunction with study visits. The QOL assessment was scheduled to occur at baseline before administration of the study medication and at every clinical visit during the five years after randomization (at three months, at six months, and every six months thereafter). However, the trial was unblinded on March 31, 1998, following an interim analysis that showed a dramatic reduction in the incidence of breast cancer among the participants who received tamoxifen. The QOL follow-up was terminated at that time due to the potential loss of the control arm. In this manuscript, as in our prior analyses of QOL data from the trial (Day *et al.,* 1999, Ganz *et al.*, 1998), we use QOL data available on participants who were recruited to the trial during the first two years of the study (June 1, 1992 to May 31, 1994) as all of these women would have been expected to have 36 months of completed follow-up data at the time the study was terminated. The sample includes 11,064 women who represent 82.6% of the total accrual to the BCPT. We use only their first three years of follow-up.

3. The Effect of Tamoxifen on Depression

When the BCPT began, there was considerable concern that tamoxifen therapy might be associated with the development of depressed mood in women with breast cancer. Although Love et al. (1991) did not find such an effect when reporting symptoms associated with tamoxifen treatment in a randomized trial in postmenopausal women with breast cancer, several researchers subsequently reported results suggesting that administration of tamoxifen might lead to depression in some breast cancer patients (Cathacart *et al.,* 1993; Shariff *et al.,* 1995; Moredo *et al.,* 1994). The latter studies were relatively small (fewer than 400 patients) and none had a placebo comparison group. However, there was a potential scientific rationale for tamoxifen's association with depression. Estrogen had been shown to have a beneficial effect on mood in postmenopausal

women (Halbreich, 1997; Gregoine, *et al.,* 1996), and it was considered plausible that tamoxifen might negate these positive effects of estrogen. Thus, careful measurement of depression, including a screening instrument to identify potential cases of depression, was important in the design of the BCPT QOL study.

The primary instrument used in the BCPT to study the change in depression level over time was the CES-D, a self-administered questionnaire (20 questions) that screens for depressive symptoms over the seven days prior to administration (Radloff, 1977). A participant's score is the sum of the responses for the 20 questions and can range from 0 (no depressive symptoms) to 60 (maximum depressive symptoms). The instrument is widely used because it is easy to administer and has excellent population-based normative data (Myers and Weissman, 1980; Roberts and Vernon, 1983; Boyd *et al.,* 1982). To assess the validity of the CES-D in the BCPT sample, we compared the baseline CES-D scores of BCPT participants with ten medical history items related to mental health that had been obtained at entry to the trial (Table 1). The first three items were obtained in the context of questions about diagnosed medical problems, although we did not verify that there had been a recorded diagnosis. There is nearly a linear relationship between the number of positives from the participant's mental health history and the CES-D score ($p<0.0001$), providing considerable reassurance that the CES-D score from this study sample was highly associated with a clinical mental health history. Similarly, the association between the mean CES-D score and the three depression-related items "ever had depression," "ever took antidepressants" (either item 4 or 6 positive), or "any two years depressed or sad", showed an increasing relationship between the CES-D score and the number of positives ($p<.001$). In addition, the baseline CES-D scores were well balanced across placebo and tamoxifen treatment assignment (Table 2). Cut-off points used in the table are somewhat arbitrary, although a cut-off of 16 is commonly used as the minimum for classifying a person as depressed (Myers and Weissman, 1980; Roberts and Vernon, 1983; Boyd *et al.,* 1982) and Lyness *et al.* (1997) used the cutoff of 22 when screening for major depression.

Table 1
BCPT Participant History Mental Health Items Obtained at Entry to the BCPT

Item	% Yes
Ever had depression	15
Ever had nervous or emotional disorder	3
Ever had psychiatric problems	1
Current antidepressants	6
Current tranquilizers	16
Previous antidepressants	4
Previous tranquilizers	15
Two weeks sad, blue, depressed, disinterested	17
Any two years depressed or sad	9
Depressed or sad most of past year	5

Table 2. Baseline CES-D Scores

Score	Placebo (%)	Tamoxifen (%)

0-10	85.5	83.9
11-15	7.8	9.0
16-21	3.8	4.1
22-60	2.8	3.0

In Figure 1, we present the mean CES-D scores by visit and treatment arm during the BCPT. The observed increase of depression among participants receiving tamoxifen is slightly less than the observed increase among participants receiving placebo, although the difference is not significant (p=0.24). Thus, the increase in the depression score during the first six months of the trial does not appear to be related to the administration of tamoxifen. It is noteworthy that the dramatic increase in scores at months 3 and 6 would almost certainly have been attributed to tamoxifen had there not been a placebo arm. This illustrates the danger of trying to establish a cause-and-effect relationship in a non-randomized setting.

FIGURE 1 ABOUT HERE.

We do not know why the CES-D depression scores increased for participants on both arms of the study (placebo and tamoxifen). It is possible that symptoms of worry and depression increased due to the controversy surrounding this trial, the fear and uncertainty of taking either placebo or active agent, an increased awareness of breast cancer risk, or a concern over potential therapy side effects. Alternatively, the raised scores might be partially attributed to "nocebo effect" (Hahn, 1997): if an individual fears or believes that a side effect may occur from a medication, he or she will report it. (As will be shown later, participants receiving placebo also reported an increase in hot flashes, but not at the same significant rate as the participants on tamoxifen.) Since neither group of women knew which pill they were taking, they may have reported increased symptoms because they feared the potential medication side effects described to them as part of the consent process. A third possibility is that the baseline scores were artificially low and the subsequent increase reflected a regression to the mean. We do not believe the baseline scores are much lower than would be expected for the educated, socioeconomically advantaged population in the trial. However, to the extent that the scores were artificially low, it could be either that women were less likely to enter the trial when they were experiencing depressive symptoms, or that they would under-report for fear of jeopardizing their inclusion in the trial. In any case, the phenomenon of an early increase in depressive symptoms appears to be independent of tamoxifen use.

However, we were concerned that there might be a treatment effect in the subset of subjects at higher risk of depression. Because 93% of the participants had baseline CES-D scores <16, and 85% had scores <11, such an effect might not be apparent in an analysis based on the entire population. To explore this possibility we divided the women into four groups of risk: zero, one to two, three to five, and six to ten "yes" responses to the mental health items listed in Table 1. There was no difference observed between tamoxifen and placebo participants in any of the four groups. Results were similar when the baseline CES-D score was used to create risk groups (CES-D scores from 0 to 11; 12 to 15; 16 to 21; or 22 or more). There was a suggestion that tamoxifen is beneficial in the high-risk group (p=0.04), although this is likely to be a statistical artifact.

The problem of missing data is common in clinical trials that assess QOL (Bernhard *et al.*, 1998a). In the BCPT, this was exacerbated by the fact that the clinical centers were not required to collect QOL data when a participant went off the study medication. As will be seen, this led to a substantial problem of non-random missing data. Only 82 participants did not fill out the CES-D form at entry (an extremely low rate of missing baseline data), and these participants were excluded from subsequent analyses. (Questionnaires that were partially completed are

considered missing in this report.) However, of the possible 76,874 post-entry forms that 10,982 remaining participants were expected to submit during the three-year period, 13,752 (18%) were missing. At the end of the third year, slightly more than 30% were missing and participants who received tamoxifen were more likely to have missing data (33% versus 30% missing, $p<0.001$). The first three rows in Table 3 present the number and percent of missing forms preceded by a protocol-specified event (such as second primary cancer, deep-vein thrombosis, ischemic heart disease, or death); missing forms preceded by early termination of therapy for a reason not specified by the protocol; and missing forms preceded by consent withdrawal by the participant; the fourth row of this table shows the number of forms that were missing when the participant was still receiving therapy. Figure 2 displays the percent of missing forms in four groups based on baseline CES-D scores. Participants who began with an elevated CES-D score were more likely to have missing data ($p<0.001$ at three years) .

INSERT FIGURE 2 AND TABLE 3 ABOUT HERE.

The average of the CES-D scores immediately preceding a missing score was higher than the average of the CES-D scores immediately preceding an observed score (Table 4), which raised the possibility that missing scores would have been higher than concurrently observed scores. The differences were almost identical in the tamoxifen and placebo arms, indicating that while the missing data might result in an underestimate of depression, the bias would be the same in both arms. When we considered other functions of preceding scores, we found that none had a stronger association with missing scores than did the immediately preceding score. In particular, the slopes between two scores preceding a missing score were not significantly different from slopes between two scores preceding an observed score. Therefore we considered some simple imputations based on the scores immediately preceding the missing scores.

INSERT TABLE 4 ABOUT HERE.

In discussing the imputation methods, we will use the following notation. The baseline and seven post-entry CES-D scores for the *j*th individual participant will be represented by the vector $\underset{\sim}{x}_j = (x_{0j}, x_{1j}, x_{2j}, x_{3j}, x_{4j}, x_{5j}, x_{6j}, x_{7j})$, where "missing" is a possible value for the CES-D score. Let $\overline{x}_{i.}^{T}$ ($\overline{x}_{i.}^{P}$) be the average CES-D among tamoxifen (placebo) participants with an observed CES-D score at the *i*th visit. We define a new set of vectors by $\underset{\sim}{x}_i^I = (x_{0i}, x_{1i}^I, x_{2i}^I, x_{3i}^I, x_{4i}^I, x_{5i}^I, x_{6i}^I, x_{7i}^I)$, where $x_{ij}^I = x_{ij}$ if x_{ij} is observed. If x_{ij} is missing, $x_{ij}^I = x_{(i-1)j}^I + \overline{x}_{i.}^{T} - \overline{x}_{(i-1).}^{T}$ for a tamoxifen participant and $x_{ij}^I = x_{(i-1)j}^I + \overline{x}_{i.}^{P} - \overline{x}_{(i-1).}^{P}$ for a placebo participant, where the imputation begins with x_{1i}^I then x_{2i}^I and so forth. The mean CES-D curves are slightly higher than in Figure 1 (where no imputation is involved), but the differences between the two curves remain nearly identical to the differences seen in Figure 1.

Although Table 4 suggests that the imputed values defined above would be appropriate for replacing missing values, we cannot rule out the possibility that the missing values mask a greater increase in depression for tamoxifen participants than for placebo participants. For example, there might have been a subset of tamoxifen participants who became depressed as a result of the treatment and dropped out before this effect could be observed. We do not have data available to verify that this is not the case. In order to see just how great a differential (by treatment) would

have been required to change the interpretation of the data, we performed three sensitivity analyses.

For the first sensitivity analysis, we imputed missing values as defined above, but for every missing value of a tamoxifen participant we added 0.5 units to the imputed value. The resultant mean values of CES-D at each assessment are almost the same between treatment arms. This is somewhat reassuring, since adding .5 units to each missing CES-D score for tamoxifen participants and none for placebo participants is extreme. As Table 3 indicated, the status of the participants with missing forms was similar on both arms. In instances in which institutions reported the reason participants went off study, only 3% reported depression as the reason for doing so.

The second sensitivity analysis was based on a partitioning of missing questionnaires into those that were missing for a variety of non-treatment-related reasons and those that were missing for treatment-related reasons. Specifically, we assumed that if m questionnaires were missing (at a particular assessment time) in the placebo arm, and $m+x$ questionnaires were missing in the tamoxifen arm, then some fraction of the x questionnaires might be attributable to excess depression caused by tamoxifen. We calculated treatment group means (at each assessment time) as if some fraction r (for various candidate values of r) of the missing tamoxifen scores were replaced with the mean of all observations at that assessment that were at least 16, since these represent severe depressive symptoms. The remaining missing values in both arms were replaced with the mean of all observations at that assessment. At $r=1/2$, the curves of imputed CES-D for the two treatment groups overlapped [not shown]. That is, there did not appear to be a tamoxifen-related increase in CES-D unless greater than half of the excess missing questionnaires were assumed to coincide with severe depressive symptoms.

All of the analyses shown above were also carried out for a binary outcome of severe depressive symptoms, defined as any CES-D score $\geq$16. In Figure 3.A, we plot the proportion of values classified as a "yes" as a function of time and again find no tamoxifen effect. Imputation of the missing values using preceding scores had minimal impact on our findings. For a sensitivity analysis, we performed the imputation with the additional assumption that 3.2% of the missing tamoxifen CES-D forms had a score $\geq$ 16, even though the prior score was <16. This would be roughly equivalent to assuming that all the tamoxifen participants who reported depression before dropping out of the study subsequently had a score $\geq$ 16, while none of the placebo participants reporting depression before dropping out had a score exceeding 15. The sensitivity analysis, presented in Figure 3.B, indicates that under this fairly extreme assumption about the drop-outs, the two curves would essentially overlap.

INSERT FIGURES 3A AND 3B HERE.

As a final step in the sensitivity analysis, we considered a model-based method that adjusts for drop-out related to observed and unobserved CES-D outcomes through subject-level random effects. This approach, which may be used to adjust for other covariates, has been presented previously in other randomized trial contexts for continuous data (Schlucter, 1992; DeGruttola and Tu, 1994) and for binary data (Ten Have *et al.,* 1998), and in a cohort study context for ordinal data (Ten Have *et al.,* 2000). More specifically, we fitted an ordinal logistic model with random effects to the CES-D outcome data. The CES-D score was categorized as in Table 2. The models make the proportional odds assumption, that is, the odds ratio specified for a given cut-point of the ordinal CES-D scale is the same as the odds ratio specified for every other cut-point. This approach is not designed for intermittent missingness. Therefore, any participant's data subsequent to a missing form was deleted for the purpose of this analysis. The model comprised three components consisting of different covariate effects but sharing the same subject-level random effect structure. The first was an ordinal CES-D outcome component with treatment arm and time (7 degrees of freedom) as main effects, and their interaction (7 degrees of freedom). The

second and third model components corresponded to separate discrete survival time logistic specifications for non-protocol and protocol specified drop-out. Each of these drop-out components included main effect covariates corresponding to treatment arm and time.

We present results based on two versions of each of these drop-out components. The first version includes as covariates the CES-D outcome before drop-out and its interaction with treatment arm and type of dropout (protocol vs. non-protocol). In the second version, each of these drop-out components excludes the CES-D outcome and its interactions. The ensuing results are based on these model specifications without baseline covariates. Including baseline age did not alter the results. The subject-level random effect structure shared by the CES-D and drop-out components induces a relationship between the CES-D observed and unobserved outcomes and the risk of drop-out. The magnitude of this relationship is characterized by the specification of separate variance components of the random effect for each of the three components in the model. Separate large variance components for the outcome component and for a drop-out component indicate a strong relationship between outcome and the respective form of drop-out. For comparison, we also present results based on the assumption that drop-out is *missing at random (MAR)*. That is, drop-out is conditionally independent of the unobserved CES-D outcomes, conditioned on all observed data (Little, 1995). In summary, we have used these three models: 1) the random effects logistic model without a drop-out component, under the assumption that drop-out is missing at random (naïve model); 2) the random effects logistic model augmented with a discrete time survival logistic model for drop-out, which shares a random effect with the ordinal CES-D outcome (Joint 1 model); and 3) model #2 with the last observed CES-D outcome added as a covariate (Joint 2 model).

The likelihood ratio test of treatment arm differences in change across time (7 degrees of freedom) was not significant (p=.14). As Table 5 suggests, this result was robust with respect to the drop-out assumptions (e.g., MAR). More specifically, the estimates of the log treatment odds ratio at baseline and corresponding treatment-time interaction terms at each follow-up time differ very little across the three models. To evaluate the strength of the relationship between outcome and drop-out, we present the variance components of the random effect shared by the three components (outcome, two drop-out types: non-protocol- and protocol-defined) two of the models, Joint 1 and Joint, 2 in Table 6. Note that the naive model only has the outcome component and therefore only one variance component. Table 6 shows that neither of the drop-out components in Joint 1 and Joint 2 models is related to the outcome through a random effect. This lack of relationship between outcome and drop-out is consistent with the fact that the log odds ratio estimates in Table 5 are very stable between the naïve and joint models. This suggests that the naïve random effects model accommodates the relationship between outcome and protocol-defined drop-out. That is, the MAR relationship under the naïve model characterizes the type of relationship between drop-out and outcome represented by the joint models. Of course, it may be that a different relationship exists that is not characterized by either the joint or naïve models.

INSERT TABLES 5 AND 6 ABOUT HERE.

In summary, our study data indicate that tamoxifen does not influence depressive symptoms among women who are at high risk for breast cancer, and there is no indication that missing data masked an effect. It appeared that the missing data did result in slight underestimates of the CES-D scores, which were increased following imputation.

4. Strategies for the Evaluation of Missing Data: Hot Flashes

Although tamoxifen did not appear to influence the CES-D score in this study, it clearly was associated with other symptoms. Numerous studies have shown that tamoxifen increased the number and severity of hot flashes in women being treated for cancer, and this effect was also seen in the high-risk women participating in the BCPT (Day *et al.*, 1999). Hot flash was the most commonly reported symptom on either arm of the BCPT.

In Figure 4 (solid lines), we present the score reported by these women for hot flashes at each cycle by treatment (possible values ranged from 0=none to 4=extreme). There is a clear increase in this symptom associated with tamoxifen throughout the study. (Note that participants taking placebo also report an increase in mean hot flash score, although this increase is not as great as for those taking tamoxifen. This may be another example of the nocebo effect.) Differences in hot flashes due to treatment are highly significant ($p<.001$) at every visit. However, when hot flash scores immediately preceding a missing value were compared to the scores immediately preceding an observed value (Table 7), there was a differential effect according to treatment. We again did an imputation in which missing values were replaced by the prior score adjusted for the mean for the visit (as described previously for the CES-D analyses). There is still clear evidence of a tamoxifen effect (dashed lines in Figure 4), but the values for the tamoxifen curve are slightly lower than when the missing values are omitted, while the values for the placebo curve remain nearly unchanged, indicating that we might be slightly overestimating the treatment effect if we ignore missing values. For example, the difference in average scores is .30 at three years when missing data are ignored versus .26 following the imputation.

INSERT FIGURE 4

Table 7. Average hot flash score prior to missing versus observed scores

	Missing subsequent questionnaire	Observed subsequent questionnaire
Placebo	0.87	0.77
Tamoxifen	1.12	1.16

An alternative analysis of these data based on the informative drop-out model used for the CES-D revealed a significant difference between the treatment arms with respect to change at each follow-up time ($p<.001$). As with the CES-D non-significant treatment difference, this significant result was robust with respect to drop-out assumptions under the random effects ordinal logistic model. The logistic model requires the assumption that the relationship between symptoms and drop-out risk is in the same direction in both the placebo and tamoxifen groups and over time. As Table 8 indicates, this assumption did not hold for the hot flash data. Hence, we were unable to adjust for the observed drop-out pattern to obtain valid estimates of the treatment effect.

INSERT TABLE 8 ABOUT HERE

5. Conclusions

Several points became clear in the analysis of the CES-D data. Perhaps the most important is that one would be likely to conclude that tamoxifen increased depressive symptoms if all the participants had received tamoxifen, as this would appear to be the most likely cause of the immediate increase in depressive symptoms. However, the randomization allowed us to see that the effect increase was comparable when participants received placebo, ruling out tamoxifen as the cause. The fact that the prior scores associated with missing values were elevated in both arms

indicated that the degree of depressive symptoms might have been underestimated slightly on both arms. However the elevation was the same in both arms, which made it unlikely that there was a differential drop-out effect by treatment. This partially explains why imputation analyses still led to the conclusion that tamoxifen did not result in increased depressive symptoms. Sensitivity analyses indicated that even if there were a fairly substantial treatment related difference in the depressive symptoms among the drop-outs, accounting for this differential effect would not change the conclusion that the depressive symptoms were not treatment related.

The situation was slightly different for the hot flash outcome. There was a clear substantial effect of tamoxifen on the incidence and severity of hot flashes. Furthermore, there was evidence of a differential drop-out effect by treatment. Imputation indicated that this resulted in a small overestimate of treatment effect. The rather unusual relationship between drop-outs and treatment presented in Table 8 would require fairly flexible models if one were to estimate and make inference regarding the effect. In future methodology studies, we will address ways to handle this drop-out pattern.

ACKNOWLEGEMENTS

We would like to give special thanks to the participants in the BCPT and the data coordinators at the participating sites, without whom this QOL study would not have been possible. We also wish to note that this work was supported in part by NIH/NCI Grant U10 CA69974 and DAMD grant 17-97-1-7058. We thank Barbara C. Good, Ph.D., for excellent editorial comments. In addition, we thank Maria Harper, Ph.D., and Ginny Mehalik, M.A., for editorial assistance with the manuscript.

REFERENCES

Bernhard, J., Cella, D.F., Coates, A.S., et al. (1998a), "Missing Quality of Life Data in Cancer Clinical Trials: Serious Problems and Challenges," *Statistics in Medicine*, 17, 517-532.

Bernhard, J. and Gelber, R.D., editors (1998b), "Workshop on Missing Data in Quality of Life Research in Clinical Trials: Practical and Methodological Issues," *Statistics in Medicine*, 17, 511-796.

Boyd, J.H., Weissman, M.M., Thompson, W.D., and Myers, J.K. (1982), "Screening for Depression in a Community Sample," *Archives of General Psychiatry*, 39, 1195-1200.

Cathacart, C.K., Jones, S.E., Pumroy, C.S., et al. (1993), "Clinical Recognition and Management of Depression in Node Negative Breast Cancer Patients Treated with Tamoxifen," *Breast Cancer Research and Treatment*, 27, 277-281.

Cella, D.F., and Bonomi, A.E. (1995), "Measuring Quality of Life: 1995 Update," *Oncology*, 9 (11-suppl), 47-60.

Day, R., Ganz, P.A., Costantino, J.P., et al. (1999), "Health-Related Quality of Life and Tamoxifen in Breast Cancer Prevention: A Report from the National Surgical Adjuvant Breast and Bowel Project P-1 Study," *Journal of Clinical Oncology*, 17, 2659-2669.

DeGruttola, V., and Tu, X.M. (1994), "Modelling Progression of CD4-Lymphocyte Count and its Relationship to Survival Time," *Biometrics*, 50, 1003-1014.

Fisher, B., Costantino, J.P., Wickerham, D.L., et al. (1998), "Tamoxifen for Prevention of Breast Cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study," *Journal of the National Cancer Institute*, 90, 1371-1388.

Gail, M.H., Costantino, J.P., Bryant, J., et al. (1999), "Weighing the Risks and Benefits of Tamoxifen Treatment for Preventing Breast Cancer," *Journal of the National Cancer Institute*, 91, 1829-1846.

Ganz, P.A. (1994a), "Long Range Impact of Clinical Trial Interventions on Quality of Life," *Cancer*, 74 (9-suppl), 2620-2624.

Ganz, P.A. (1994b), "Quality of Life Measures in Cancer Chemotherapy: Methodology and Implications," *PharmacoEconomics*, 5, 376-388.

Ganz, P.A., Day, R., Ware, J.E., et al. (1995). "Base-Line Quality-of-Life Assessment in the National Surgical Adjuvant Breast and Bowel Project Breast Cancer Prevention Trial," *Journal of the National Cancer Institute*, 87, 1372-1382.

Ganz, P.A., Day, R., Costantino, J. (1998), "Compliance with Quality-of-Life Data Collection in the National Surgical Adjuvant Breast and Bowel Project (NSABP) Breast Cancer Prevention Trial," *Statistics in Medicine*, 17, 613-622.

Gregoine A.J., Kumar, R., Everitt, B., et al. (1996), "Transdermal Oestrogen for Treatment of Severe Postnatal Depression," *Lancet*, 347, 930-933.

Hahn, R. A. (1997). "The Nocebo Phenomenon: Concept, Evidence, and Implications for Public Health"

Halbreich, U. (1997). "Role of Estrogen in Postmenopausal Depression," *Neurology*, 48, 1718-1729.

Little, R.J.A. (1995). "Modeling the Drop-Out Mechanism in Repeated Measures Studies," *Journal of the American Statistical Association*, 90, 1112-1121.

Love, R.R.L., Cameron, L., Connell, B.L., et al. (1991), "Symptoms Associated with Tamoxifen Treatment in Postmenopausal Women," *Archives of Internal Medicine*, 151, 1842-1847.

Lyness, J.M., Noel, T.M., Cox, C., et al. (1997), "Screening for Depression in Elderly Primary Care Patients: A Comparison of the Center for Epidemiologic Studies-Depression Scale and the Geriatric Depression Scale," *Archives of Internal Medicine*, 157, 449-454.

Moinpour, C.M., Savage, M.J., Troxel, A., et al. (1998), "Quality of Life in Advanced Prostate Cancer: Results of a Randomized Therapeutic Trial," *Journal of the National Cancer Institute*, 90, 1537-1544.

Moredo, A.T., Anelli, A., Tran, K.N., et al. (1994), "Tamoxifen Administration is Associated with a High Rate of Treatment Limiting Symptoms in Male Breast Cancer Patients," *Cancer*, 74, 74-77.

Myers, J.K., and Weissman, M.M. (1980). "Use of a Self-Report Symptom Scale to Detect Depression in a Community Sample," *American Journal of Psychiatry*, 137, 1081-1084.

Radloff, L.S. (1977), "The CES-D Scale: A Self-Report Depression Scale for Research in the General Population," *Applied Psychological Measurement*, 3, 385-401.

Roberts, R.E., and Vernon, S.W. (1983), "The Center for Epidemiologic Studies Depression Scale: Its Use in a Community Sample," *American Journal of Psychiatry*, 140, 41-46.

Schluchter, M. (1992), "Methods for the Analysis of Informatively Censored Longitudinal Data," *Statistics in Medicine*, 11, 1861-1870.

Shariff, S., Cumming, C.E., Lees, A., et al. (1995), "Mood Disorder in Women with Early Breast Cancer Taking Tamoxifen, an Estradiol Receptor Antagonist: An Unexpected Effect?" *Annals of the New York Academy of Sciences*, 761, 365-368.

Sherbourne, C.D. (1992), In: Stewart A.L., Ware, J.E. Editors. *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach,* Durham (NC): Duke University Press, 194-204.

Tannock, I.F., Osoba, D., Stockler, M.R., et al. (1996), "Chemotherapy with Mitoxantrone plus Prednisone or Prednisone Alone for Symptomatic Hormone-Resistant Prostate Cancer: A Canadian Randomized Trial with Palliative End Points," *Journal of Clinical Oncology*, 14, 1756-1764.

Ten Have, T.R., Miller, M.E., Reboussin, B.A., et al. (2000), "Mixed Effects Logistic Regression Models for Longitudinal Ordinal Functional Response Data with Multiple Cause Drop-Out from the Longitudinal Study of Aging," *Biometrics*, 56, 279-287.

Ten Have, T.R., Pulkstenis, E., Kunselman, A., et al. (1998), "Mixed Effects Logistic Regression Models for Longitudinal Binary Response Data with Informative Drop-out," *Biometrics*, 54, 367-383.

Ware, J.E., Kosinski, M., and Keller, S.D. (1994), *SF-36 Physical and Mental Summary Scales: A User's Manual.* Boston, MA: The Health Institute, New England Medical Center.

201 N. CRAIG STREET, SUITE 350
NSABP BIOSTATISTICAL CENTER
PITTSBURGH, PA 15213
UNITED STATES

Figure Legends:

Fig. 1. Change from baseline score for depression in participants in the BCPT. Depression is slightly increased in the placebo group, compared to the tamoxifen group (not statistically significant).

Fig. 2. The percent of missing questionnaires at each visit by baseline CES-D group (0-10, 11-15, 16-21, and 22-60) which is higher for subjects with higher baseline CES-D scores.

Fig. 3. A. Increase in percent of participants whose CES-D score was at least 16, minus the percent at baseline. The percent increased in both arms.

B. Effect of missing data. Increase in percent of participants whose CES-D score was at least 16, after imputation with the previous observed score, adjusting for the difference in treatment arm means between the missed visit and the preceding visit. The imputed observations in the tamoxifen arm had an additional 3.2% added, and the resulting curves are nearly overlapping.

Fig. 4. The mean hot flash score after subtraction of each participant's baseline score, by treatment arm (solid lines) and the mean hot flash score after subtraction of each participant's baseline score, by treatment arm (dashed lines). Tamoxifen subjects experienced more severe hot flashes. For each subject, missing values were first imputed with previous observed values, adjusting for the difference in treatment arm means between the missed visit and the preceding visit. Imputation did not substantially change the comparison.

Table 3
Number and Percent of Forms Missing by Status of Patient

	Placebo	Tamoxifen
Protocol Event	801 (12%)	791 (11%)
Stopped Therapy	3347 (50%)	3883 (55%)
Withdrew Consent	1359 (20%)	1297 (18%)
On Therapy	1140 (17%)	1134 (16%)
Total	6647	7105

Table 4
Average CES-D Score Prior to Missing vs. Observed Scores

	Avg. CES-D Scores	
	Before Missing Score	Before Observed Score
Tamoxifen Arm	7.78	6.50
Placebo Arm	7.70	6.44

*A score of 16 or higher is considered an indicator of depression.

Table 5

Effect of missing: For CES-D scores, estimates of baseline treatment log odds ratio and corresponding interactions between treatment and time for each follow-up time (standard errors in parentheses) for three models[1]

Model	Base Tx LogOR	Tx by Time Interaction Log OR at Follow-up Times (months)						
		3	6	12	18	24	30	36
Naïve	0.17 (0.08)	-0.07 (0.09)	-0.11 (0.09)	-0.20 (0.09)	-0.18 (0.09)	-0.17 (0.09)	-0.26 (0.10)	-0.19 (0.10)
Joint 1	0.18 (0.08)	-0.08 (0.09)	-0.12 (0.09)	-0.21 (0.09)	-0.19 (0.09)	-0.18 (0.10)	-0.27 (0.10)	-0.20 (0.10)
Joint 2	0.17 (0.08)	-0.08 (0.09)	-0.12 (0.09)	-0.21 (0.09)	-0.19 (0.09)	-0.17 (0.10)	-0.27 (0.10)	-0.19 (0.10)

[1]Models: 1) the random effects logistic model without a drop-out component under the assumption drop-out is MAR (naïve model); 2) the random effects logistic model augmented with a discrete time survival logistic model for drop-out that shares a random effect with the ordinal symptom outcome (joint 1 model); 3) model 2) with the last observed symptom outcome added as a covariate (joint 2 model).

Table 6:

For CESD, estimates of variance components of random intercepts for three models[1]

Model	Symptom Outcome Component	Non-Protocol-Specified Drop-out	Protocol-Specified Drop-out
Naïve	5.25	NA[2]	NA[2]
Joint 1	5.91	0.05	0.01
Joint 2	5.59	0.04	0.02

[1] Models: 1) the random effects logistic model without a drop-out component under the assumption drop-out is MAR (naïve model); 2) the random effects logistic model augmented with a discrete time survival logistic model for drop-out that shares a random effect with the ordinal symptom outcome (joint 1 model); 3) model 2) with the last observed symptom outcome added as a covariate (joint 2 model).

[2] NA: not applicable because naïve model does not include drop-out components

Table 8
Difference in Mean Preceding
Missing HFS – Observed HFS: Logistic approach
would be appropriate if the values all had the same sign

Visit (Mo.)	Placebo	Tamoxifen
3	0.0259	0.0384
6	0.1218	0.1418
12	0.1212	-.0484
18	0.1037	-.0952
24	-.0272	-.0850
30	-0.0594	-.2230
36	.0862	-.1182

Figure 1

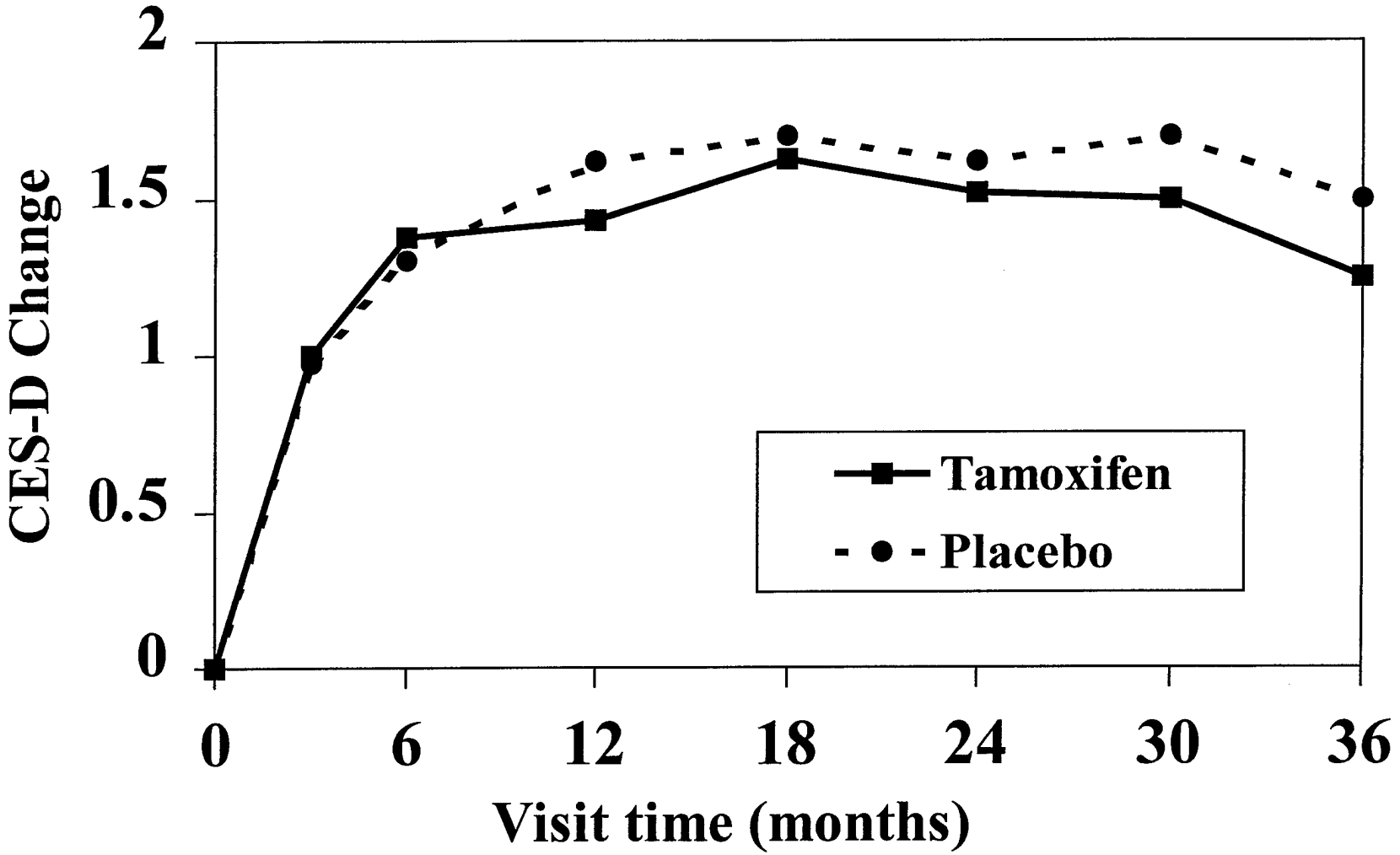

2
1.5
1
0.5
0
CES-D Change
0
6
12
18
24
30
36
Visit time (months)
Tamoxifen
Placebo

Figure 2

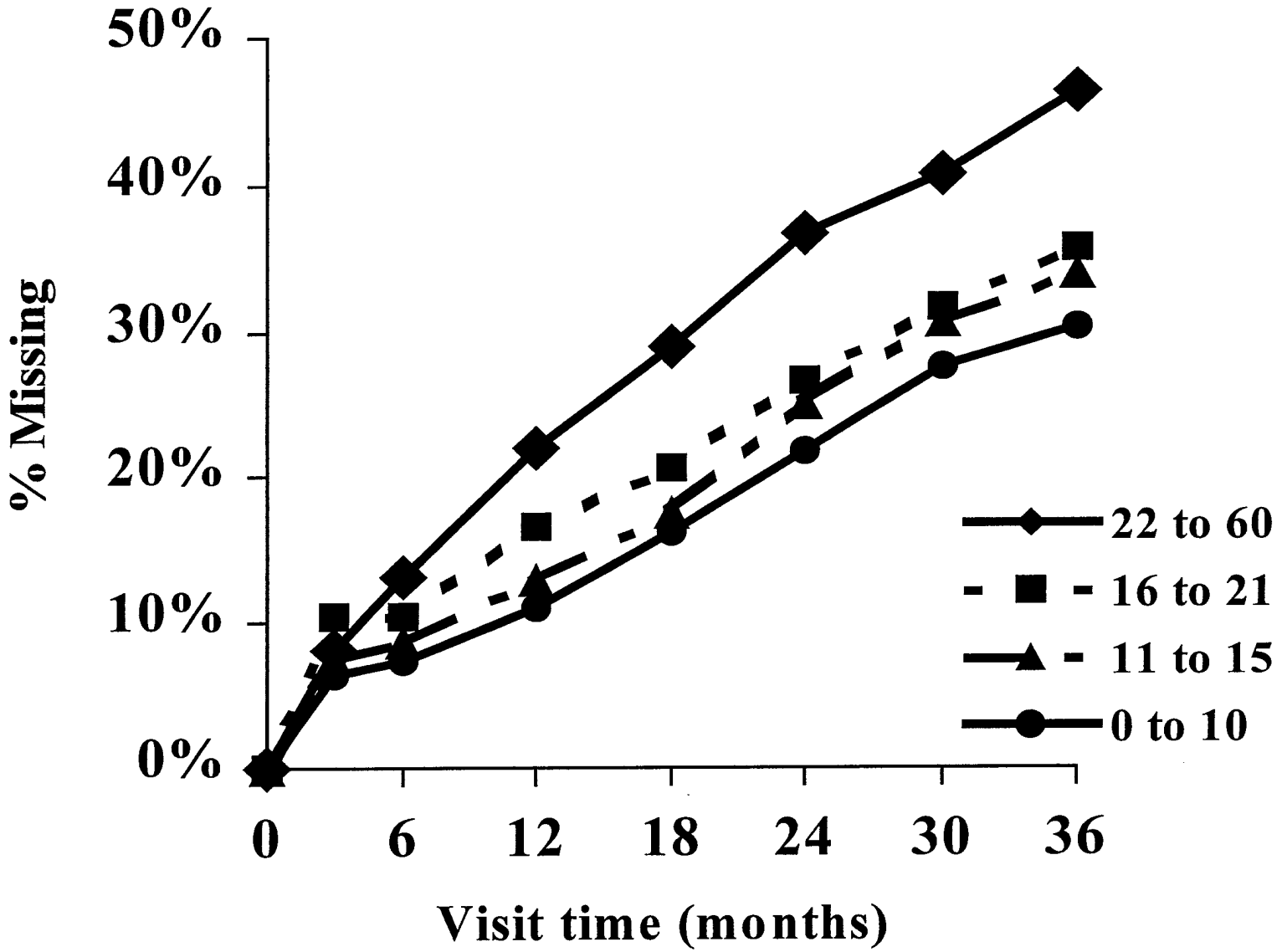

50%
40%
30%
20%
10%
0%
% Missing
0
6
12
18
24
30
36
Visit time (months)
22 to 60
16 to 21
11 to 15
0 to 10

Figure 3.A

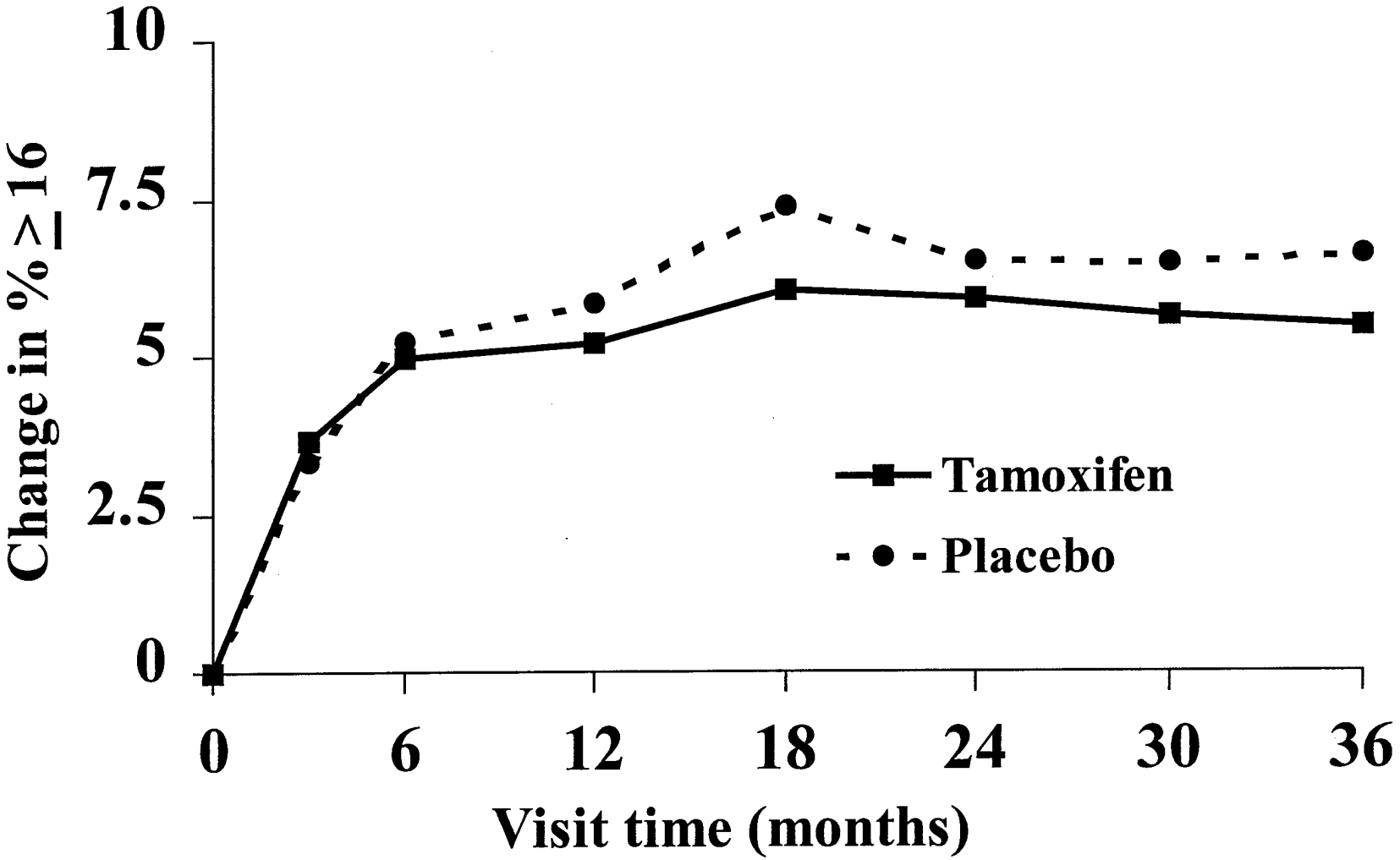

10
7.5
5
2.5
0
Change in % ≥ 16
0
6
12
18
24
30
36
Visit time (months)
Tamoxifen
Placebo

Figure 3.B

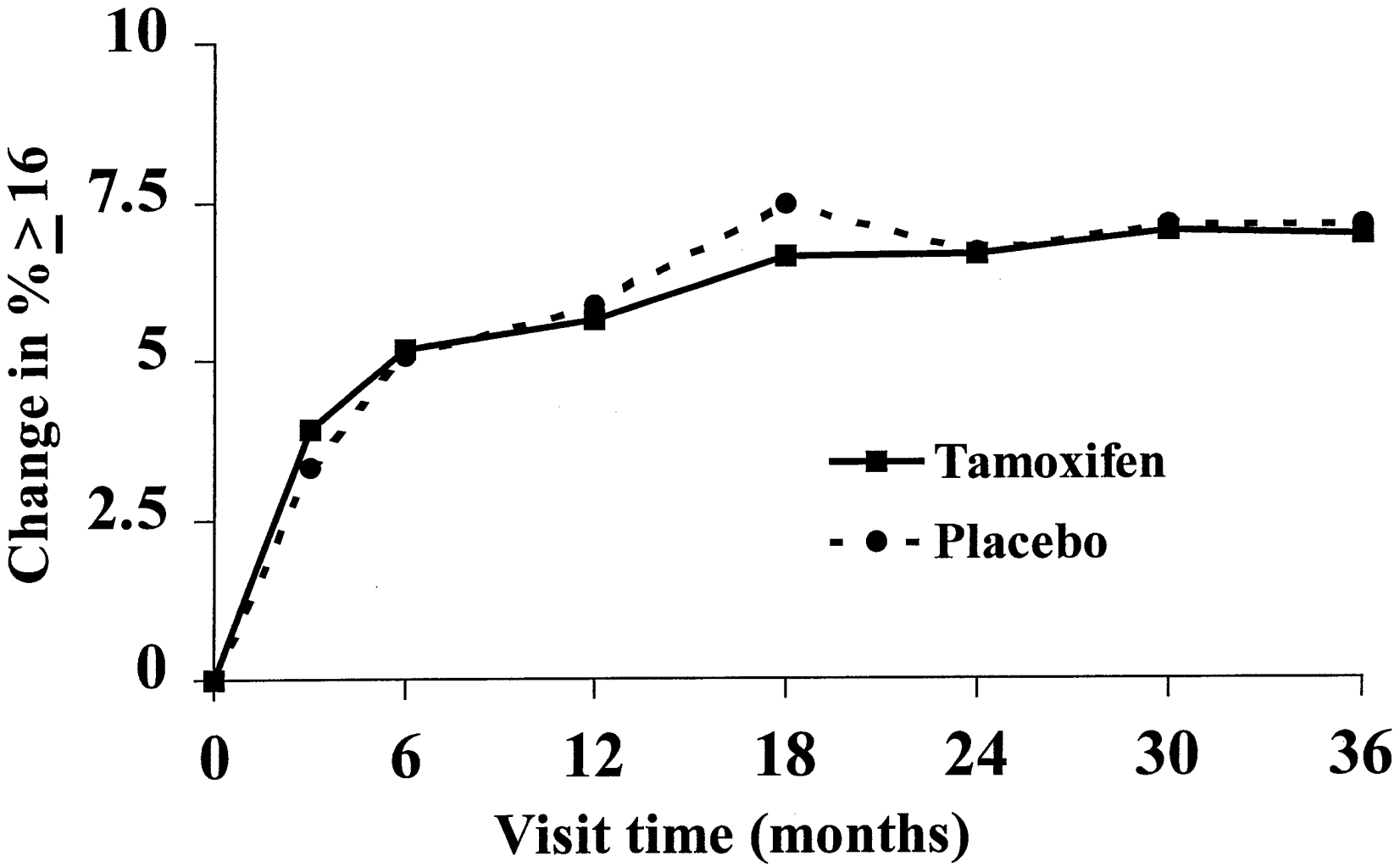
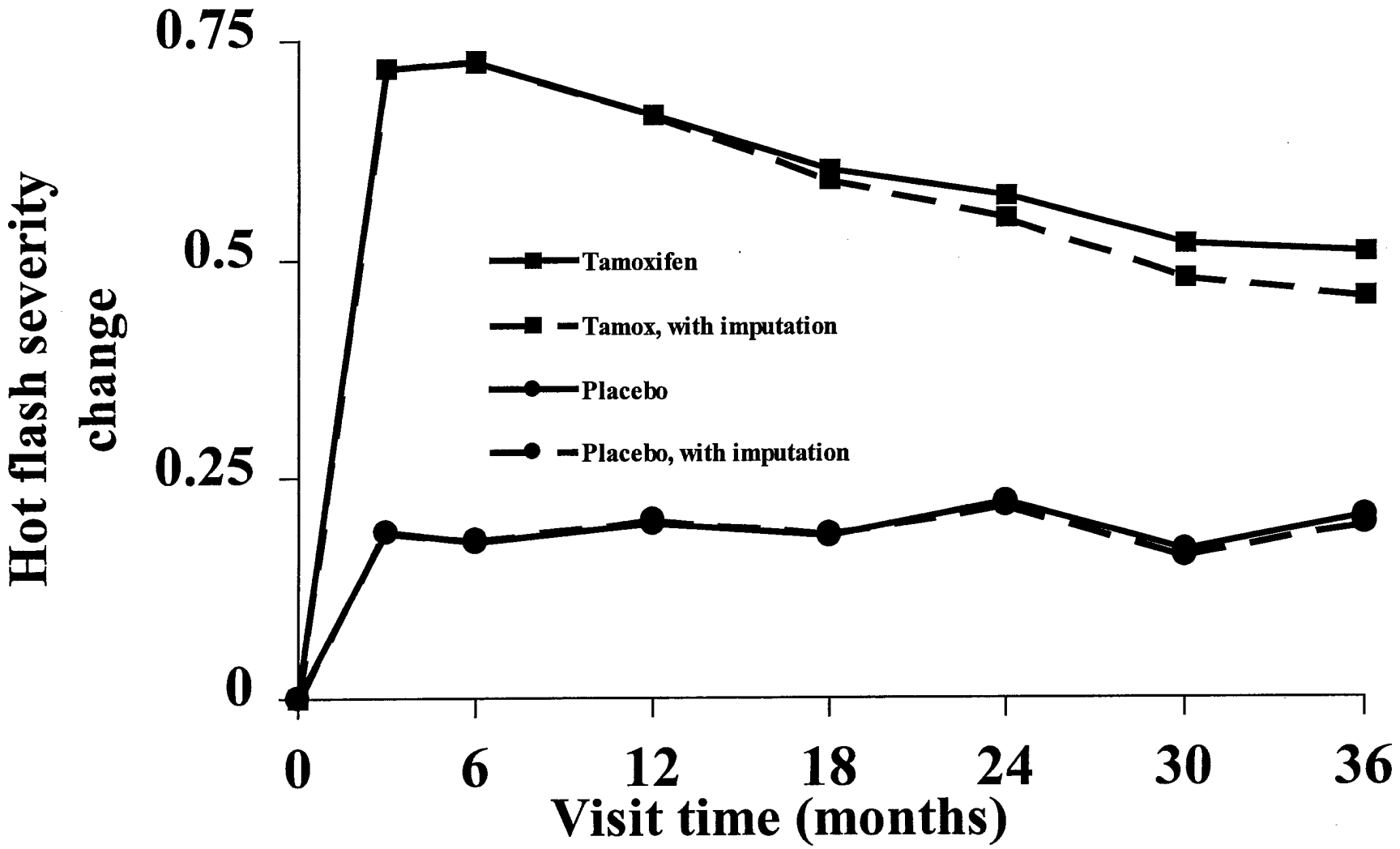

10
7.5
5
2.5
0
Change in % ≥ 16
0
6
12
18
24
30
36
Visit time (months)
Tamoxifen
Placebo

Figure 4


0.75
0.5
0.25
0
Hot flash severity change
Tamoxifen
Tamox, with imputation
Placebo
Placebo, with imputation
0
6
12
18
24
30
36
Visit time (months)

Appendix 4

Kiebert G, Wait S, Bernhard J, Bezjak A, Cella D, Day R, Houghton J, Moinpiour C, Scott C, Stephens C. Practice and policy of measuring quality of life and health economics in cancer clinical trials: a survey among cooperative groups. Quality of Life Research 2000; 9(10):1073-80.

Quality of Life Research 9: 1073–1080, 2001.

Practice and policy of measuring quality of life and health economics in cancer clinical trials: A survey among co-operative trial groups

G. Kiebert[1,2], S. Wait[2], J. Bernhard[3], A. Bezjak[4], D. Cella[5,6], R. Day[7], J. Houghton[8], C. Moinpour[9], C. Scott[10] & R. Stephens[11]

[1]*MEDTAP International, London, UK;* [2]*Quality of life Taskforce in Oncology, Novartis Pharma, Basel, Switzerland;* [3]*Swiss Group for Clinical Cancer Research, and International Breast Cancer Study Group, Bern, Switzerland;* [4]*National Cancer Institute of Canada, Clinical Trials Group, Toronto, Ontario, Canada;* [5]*Eastern Cooperative Oncology Group, Evanston, IL, USA;* [6]*Gynecologic Oncology Group, Evanston, IL, USA;* [7]*National Surgical Adjuvant Breast and Bowel Project, Pittsburgh, PA, USA;* [8]*Cancer Research Council and UCL Cancer Trials Centre, London, UK;* [9]*Southwest Oncology Group, Seattle, WA, USA;* [10]*Radiation Therapy Oncology Group, Philadelphia, PA, USA;* [11]*Medical Research Council, Clinical Trial Unit, London, UK*

Accepted in revised form 2 February 2001

Abstract

Background: Co-operative groups have played an important role in the advance of health-related quality of life (HRQL) research. However, definitions of the concept, criteria for selection of existing instruments and methods for data collection and interpretation remain poorly defined in the literature. A survey was conducted amongst the major cancer co-operative groups in order to gain a better understanding of their current policy and processes to ensure optimal HRQL data collection within cancer clinical trials. The topic of health economics was similarly addressed. *Methods*: A written questionnaire was addressed to 16 major European and North American cancer co-operative groups. Eleven groups responded (response rate: 69%), however, one group could not provide information for the survey, thus ten questionnaires were available for analysis. *Results*: The results from this survey among co-operative groups show that HRQL (more than health economics) is recognized as an important, although usually secondary, outcome measure in oncology trials. On the whole, co-operative groups have a rather flexible policy towards the inclusion of HRQL (and HE) into their clinical trials, and practice is very much on a case-by-case basis, but use standard practice guidelines and internal procedures is to ensure well-defined study protocols and enhance good quality studies.

Key words: Cancer, Co-operative group, Health economics, Randomized controlled clinical trials, Quality of life

Introduction

In chronic diseases where cure is often not achievable, it has long been recognized that improvement in health-related quality of life (HRQL) is of great importance. Oncology was one of the first disease areas where the importance of HRQL as an outcome measure was acknowledged; in US, HRQL outcomes were first included in large treatment and prevention trials in cardiovascular disease. Over the past ten years, there has been an increasing emphasis on the role of alternate outcomes other than the classical clinical trial endpoints of response rate, disease-free or overall survival. Since most trials take many years to mature, it is only now that gradually more and more publications of clinical trials include HRQL.

This study was supported from an unrestricted financial grant from Novartis Pharma AG.

Co-operative groups are playing an increasingly important role in the advancement of cancer care through the conduct of clinical trials, and the establishment of treatment recommendations and guidelines. Collaborative trials groups have also been active proponents of quality of life research. For instance, the proceedings of a workshop focusing on practical and methodological issues related to missing quality of life data in clinical trials in which all major co-operative trial groups participated and contributed were recently published as a special issue in Statistics in Medicine [1].

An informal review of existing literature indicated that many of the large oncology co-operative groups have some kind of policy or guidelines for the inclusion of HRQL as an endpoint in cancer clinical trials. However, the overall information from existing publications is scarce, incomplete and not up-to-date. In particular, information on criteria for selection of existing instruments, methods for assessment, and data collection procedures and instructions is lacking. For this reason, a survey was done of the major co-operative groups (i) that conduct clinical studies in more than one type of cancer or (ii) that focus on a single type of cancer but whose scope and membership are pan-continental.

The objective of this survey was to obtain up-to-date information on the co-operative group policy on HRQL research. Since health economics (HE), specifically resource utilization data collection, is gradually being evaluated in cancer clinical trials, our survey addressed this as an additional topic.

The survey was developed and conducted within the context of a special multidisciplinary taskforce, whose mandate was to develop internal guidelines on HRQL evaluation within oncology clinical trials at a large pharmaceutical company. Recognizing the prominent role that co-operative groups have played in HRQL research in oncology, the taskforce felt that it was essential to look to these groups for 'state of the art' processes and strategies to ensure optimal HRQL data collection within clinical trials.

Methods

The target group consisted of all major national or international co-operative groups that conduct studies in more than one type of cancer and multicontinental groups focusing on one type of cancer. The first step involved the identification of the key person in each co-operative group responsible for quality of life issues who could respond to the questionnaire on behalf of the co-operative group. This step was performed by telephone survey by the principal study investigator (GK). For all groups this key person is a specialized quality of life researcher. Once the key person was identified, this person was sent a cover letter stating the objective and content of the survey, an invitation to participate, and a request to return the completed questionnaire within six weeks. A written reminder was sent to all non-responders after six weeks. Three weeks thereafter, the remaining non-responders were contacted by telephone and, in one case, by fax.

The final response rate was 11 out of 16. Three groups did not respond, two groups refused (one because of time constraints (Cancer and Leukemia Group B (CALGB)) and one because of concerns about confidentiality of information (European Organisation for the Research and Treatment of Cancer (EORTC)). One group was willing to participate, but at the time of the survey this information was not readily available for organizational reasons. Thus, a total of ten questionnaires were available for analysis. Table 1 provides an overview of the groups that were approached and their responses to our invitation to participate in the survey.

The questionnaire was developed especially for this survey. A listing was made of all relevant topics for which we intended to collect data. In a second step a set of questions were formulated addressing all different aspects of each topic. A draft version of the survey was reviewed by members of the taskforce experienced in the development of questionnaires.

The questionnaire addressed the following topics: overview of ongoing clinical trials with and without HRQL in the most prevalent types of cancer; co-operative group trial selection policy; procedures and methods for inclusion of HRQL into clinical trials; study center training and guidelines for HRQL data collection; data analysis and reporting of findings. The same questions were asked for HE. The results of the survey are discussed below in this order of topics.

Results

Overview of ongoing clinical trials

Most numerous of on-going clinical trials are those in gynecological, breast, lung, prostate and colorectal cancers (Table 2). In more than half of these trials, HRQL is evaluated, although usually as a secondary endpoint, and only seldom as the primary endpoint. Notable exceptions are trials that evaluate best supportive care, where HRQL is the primary endpoint in six out of eight trials. HE endpoints are much less frequently collected in the reported trials.

Trial selection policy

Limited research resources and budget constraints often necessitate prioritising of HRQL studies. In the context of clinical trials this situation is not

Table 1. Overview of target groups and survey response

Co-operative group	Response
Cancer Research Council, UCL Cancer Trials Centre (UK)	Yes
Medical Research Council, Clinical Trials Unit (UK)	Yes
Swiss Group for Clinical Cancer Research (SIAKK/SAKK) (Switzerland)	Yes
International Breast Cancer Study Group (Switzerland)	Yes
National Cancer Institute of Canada, Clinical Trials Group (Canada)	Yes
Eastern Cooperative Oncology Group (US)	Yes
Gynecologic Oncology Group (US)	Yes
Southwest Oncology Group (US)	Yes
National Surgical Adjuvant Breast and Bowel Project (US)	Yes
Radiation Therapy Oncology Group (US)	Yes
Deutsche Krebsgesellschaft (as representative of the German Co-operative Groups (Germany))	Willing, but information not readily available
Interdisciplinary Group for Cancer Care Evaluation (Italy)	No response
Fédération Nationale des Centres de Lutte Contre le Cancer (France)	No response
Nordic Cancer Trial Group (Scandinavia)	No response
Cancer and Leukemia Group B (US)	Refused
European Organisation for the Research and Treatment of Cancer (Europe)	Refused

Table 2. Summary of ongoing clinical trials by disease site in 1998 for all ten groups surveyed

Cancer site*	Total number of ongoing trials in 1998 (n)	Number of trials with HRQL as primary endpoint n (%)	Number of trials with HRQL as secondary endpoint n (%)	Number of trials including HE n (%)
Brain	11	0	5 (45)	0
Breast	30	3 (10)	14 (47)	1 (3)
Colorectal	20	1 (5)	9 (45)	3 (15)
Prostate	19	3 (16)	10 (53)	0
Gynecology	32	1 (3)	16 (50)	5 (16)
Head & Neck	16	0	6 (38)	0
Leukemia	10	0	0	0
Lung	27	0	13 (48)	2 (7)
Lymphoma	11	1 (9)	1 (9)	0
Melanoma	3	0	1 (33)	0
Multiple cancer sites				
Supportive care	9	6 (67)	2 (22)	0
Palliative care	9	2 (22)	6 (67)	2 (22)

* Selection of type of cancer based on prevalence of the disease. It does not represent a complete overview of all ongoing clinical trials per group.

different. The trials where HRQL is considered most important are those in which a large survival advantage is not expected; which compare very different treatments (e.g., chemotherapy vs. radiation) that will likely result in different side effect profiles, and those in which patients are symptomatic and the treatment is expected to relieve those symptoms.

In nine out of ten groups, there is a specific person or committee in the co-operative group responsible for HRQL research issues such as trial selection, procedures for data collection, implementation, and methodology. Only one of the co-operative groups has adopted a policy of including HRQL in all cancer clinical trials as a standard (National Cancer Institute of Canada (NCIC)). In all other groups, this decision depends on a number of factors such as study design, research question, sample size, number of participating centres and countries, and a number of population characteristics. A randomized study design allows for comparison of HRQL between the two study arms and distinguishes the effect of trial intervention over time. The research question determines the relevance of HRQL as an endpoint to that question and the sample size distinguishes whether there will be a sufficient number of patients to provide an answer to the HRQL research question. The number of participating centres and countries influence the feasibility of HRQL assessment and likelihood of compliance to questionnaire completion, the number of languages in which the questionnaire will need to be available, as well as funding needed. Duration of the trial affects feasibility as well as funding issues. Financial constraints can play limitating role and necessitate prioritising of trials that include HRQL as an outcome parameter. Age of the patients is most relevant in the paediatric population to determine whether self-assessment of HRQL is possible. And lastly, the health care setting frequently influences the availability of personnel to administer HRQL questionnaires.

Table 3 provides the detailed ratings of importance of different factors in the selection of trials for inclusion of HRQL. Numbers represent the sum of responses from the 10 groups surveyed. Globally, treatment characteristics appear to play a more important role in the selection of trials for HRQL data evaluation than trial and population characteristics. Study design, available resources, toxicity of treatment and absence of incremental survival advantage were the most important factors.

There is often discussion as to whether HRQL is best collected within the actual clinical trial or as a separate or companion protocol. When asked whether HRQL studies were conducted as an integral part of the study protocol, seven groups responded 'yes, always', and three reported 'sometimes'. Six respondents stated that HRQL was never conducted with a separate protocol, and four respondents stated that this was sometimes the case. To the question whether, when included in a trial, HRQL was a mandatory aspect of the study for all participating centers, five groups responded 'yes, always', one 'no, never' and four 'sometimes'.

Mode of assessment and choice of instrument

All but one group use written questionnaires as a standard mode of HRQL assessment, and five groups use in principle the same instrument in all studies (either EORTC QLQ-C30; FACT-G; or LASA scale). For the other groups, the choice of the instrument depends mainly on the trial characteristics, psychometric properties and its practicality for a particular trial, and to a lesser degree on language availability, familiarity with the instrument or its theoretical foundation. Examples of questionnaires that have been used previously in trials are SWOG QoL questionnaire, CARES-SF, MOS-SF36, EORTC QLQ-C30; FACT-G; or LASA scale.

HRQL research guidelines

All groups provide some form of specific instructions to the participating centers for the collection of HRQL data. These can consist of written guidelines, training days, a HRQL training video, procedure manuals for HRQL assessment, regular internal training at group meetings, and an initiation site visit to discuss the HRQL aspects of the protocol.

Six out of ten groups have written internal procedures or guidelines for HRQL data analysis and interpretation. Topics covered by all guidelines include the plan for statistical data analysis

Table 3. Average importance of factors influencing decisions to include HRQL as an endpoint in a clinical trial

	Importance rating			
	Not at all	A bit	Quite a lot	Very much
Trial characteristics				
Resources available	0	2	4	4
Study design	2	2	2	4
Monitoring capacity	2	2	5	1
Representativeness of participating investigators and centers	3	4	1	2
Sample size	1	6	3	0
Participating countries	1	8	0	0
Number of participating countries	2	7	0	0
Number of centers	4	6	0	0
Duration of trial	5	4	1	0
Treatment of characteristics				
Equal efficacy in terms of survival expected	0	0	2	7
Toxicity of treatment	0	0	6	3
New treatment modality	0	1	6	1
New mode of administration	0	5	3	1
Palliative intent	2	1	2	3
Curative intent	2	4	2	0
Population characteristics				
Age (children, adults, elderly)	2	6	1	1
Representativeness of trial population	3	4	3	0
Health care setting (in- vs. outpatient department or home care)	2	8	0	0
Instrument characteristics				
Availability of suitable instrument	0	6	1	3
Other				
Burden on patients				
Statement that HRQL outcome is critical for interpreting results				
Potential outside funding				

Note: Answers shown above represent the sum of respondents choosing that category.

and calculation of sample size estimations. Handling of missing data is included in five out of six. Other topics mentioned were the interpretation of results as clinically meaningful changes over time (n = 1), in relation to clinical data (n = 3) or to other outcome measures (n = 1). Only one group addresses the issue of the pooling of data for multinational analysis, which is not surprising as the majority of respondents are groups that operate mainly at a national level.

Topics that are not addressed at all in existing guidelines are the dissemination of results within clinical practice and the role of HRQL outcomes in subsequent treatment decision making.

Interest in HRQL research

Four groups stated that their interest in HRQL research is very high, and five groups expressed quite some interest (missing n = 1).

Health economics

In general, the activity and interest in health economics is significantly less among all groups than for HRQL. In three groups, health economics data in the form of resource use such as hospitalization, medication, diagnostic tests used, number of outpatient visits, have never been assessed in any trial.

Four groups have a person or committee specifically responsible for health economic issues; one group has a broad outcomes committee that can address health outcomes including HE. None of the groups has a standard policy to collect HE data in each trial.

Four groups identified formal criteria that they followed when deciding whether to include HE as an outcome measure. The most important considerations were the direct cost of the investigated treatment(s), costs associated with treatment of adverse events, and potential financial consequences of treatment for the hospital, practice, or patient. Trial population characteristics and external requirements from health authorities and/or medical ethics committees play a less important role in HE inclusion decisions.

Three groups have some sort of guidelines for the collection of HE data. None of the groups has internal procedures or guidelines for the analysis and interpretation of HE data.

Interest in HE research

The perceived level if interest in HE is fairly low: one group is very interested, two groups are quite interested and five groups indicated a bit of interest in the subject (missing: n = 1).

Discussion

The objective of this study was to obtain up-to-date information of the processes and strategies used by large national and international oncology co-operative groups to conduct HRQL research and to ensure optimal HRQL data collection within their clinical trials. Questions were also asked with regards to the groups' policy towards HE data collection, as it is felt that this is an emerging, and complementary, field of research to that of HRQL [2].

One of the important limitations of our study is the size and representativeness of the study sample. We approached only (i) large national or international co-operative groups that conduct studies on more than one type of cancer and (ii) multi-continental groups focusing on one type of cancer. Moreover, we did not include groups active in the field of pediatric oncology. As a result, there are clear limitations regarding the representativeness of our sample and the generalizability of the results. The majority of the participating co-operative groups is North American, leaving other continents, and especially Europe, clearly under-represented. Non-participation in our survey does not imply lack of experience or policies regarding HRQL and HE research. For instance, the EORTC has been active in the field of HRQL research since many years, and has published on their strategy to include HRQL as an endpoint in their clinical trials [3]. It would be inappropriate to infer their policy from publicly available information as these will not provide the same level of detail obtained by our survey. The same approach would also have to be applied to other co-operative groups, and published reports from other multinational or national European groups on HRQL and HE policies and strategies are scarcer.

One may ask the question whether Europe is different from North America in its approach to HRQL research. One source of information is to look at the stance of health authorities to HRQL in these two continents. In US, a 1996 publication [4] on the position of the Federal Drug Administration (FDA) with regards to HRQL suggests that, for the FDA, HRQL is more important than traditional measures of efficacy such as tumor response for drugs that do not have any impact on survival. More recently, the FDA has set up a special committee in collaboration with outside researchers to investigate further the role of HRQL within the registration and labeling of oncology products (i.e. Subcommittee of the Oncology Drug Advisory Committee). In Europe, the European Medicines Evaluation Agency (EMEA) cite "symptom control backed up by quality of life assessments" as one of the possible secondary outcome measures in their 1996 publication of the Committee for Proprietary Medicinal Products (CPMP) [5]. However, the actual role that HRQL data have played in drug approval decisions by both of these agencies remains to be elucidated [6]. One positive example in the US is the role played by HRQL data, specifically reduction of pain, in the approval of mitoxantrone and prednisone for the treatment of hormone-refractory prostate cancer [7]. Indeed, it may be assumed that authorities in both Europe and North America are at the early stages of learning about the value of

HRQL research and findings to the development and acceptance of new cancer therapies. Within this learning environment, co-operative groups in all continents may play an important role in setting precedents, disseminating research findings and advancing methodologies in this growing field.

In our survey, we did not ask respondents to differentiate between trials that are financially supported publicly or by the pharmaceutical industry. Clinical trials in US are predominantly sponsored by the government, whereas co-operative groups in Europe and Canada have more of a mixture of government and industry sponsored studies. For industry sponsored trials, the most influential factor on whether to include HRQL as an endpoint is the requirement of this type of data by the regulatory authorities. From the perspective of the co-operative group, the issue of available funding is of great importance and can to a certain extent influence the support for HRQL assessments. Industry reimbursement rates per patient participating in a trial are usually greater than funding rates from public sources and the added resources can be used to pursue non-traditional endpoints or to provide financial support for studies involving non-pharmaceutical therapeutic modalities. It would be very interesting to conduct a similar survey among pharmaceutical companies and to be able to compare the pharmaceutical policies regarding the inclusion of HRQL and health economic research questions in clinical trials to those of co-operative groups.

The results from this survey among co-operative groups show that HRQL is recognized as an important, although usually secondary, outcome measure in oncology trials. Although health economics data such as hospitalizations or other resource use play a much lesser role in the clinical trial context, their role in reimbursement decisions may be more prominent. On the whole, co-operative groups have a rather flexible policy towards the inclusion of HRQL (and HE) into their clinical trials, and practice is very much on a case-by-case basis. The fact that many groups have developed written internal procedures or guidelines does not mean that they adopt a rigid approach towards design, analysis or interpretation of results. The purpose of guidelines and internal procedures is to ensure well-defined study protocols and enhance good quality studies. This is underlined by the fact that all groups recognized the importance of training of clinical trial managers for HRQL data collection, an aspect often neglected in industry-run HRQL studies. The fact that HRQL evaluation was most often recognized as an integral, and often mandatory, part of clinical trials is a promising sign, as acceptance and understanding of this outcome by treating physicians will only grow with their increased exposure to its analysis within the context of other clinical findings.

One aspect that was not addressed by all groups was the dissemination and positioning of HRQL findings within the context of clinical trial evidence and the implications of these findings for clinical practice. The need for further research and guidance in this area was also highlighted in several surveys of practicing oncologists on their perception of HRQL [8, 9]. Clearly, an essential aspect to the development of HRQL research remains the proper interpretation of findings, clear communication of the results to practicing physicians and patients, and, ultimately, the integration of HRQL aspects of therapy into actual treatment decisions.

In conclusion, the results of this survey confirm the impression that HRQL research is a growing, however still developing field in the context of clinical trials. Co-operative groups are likely to continue to play an increasing role in the advancement of this science and the dissemination of findings to treating physicians and their patients. Their role in the promotion of health economics research may be a lesser one. One may hope that the knowledge and experience that these trials groups acquire in including HRQL parameters into their trials may serve other researchers and drug sponsors in achieving a more comprehensive assessment of the impact of new therapies on cancer patients.

References

1. Statistics in Medicine 1998; 17.
2. Fitzpatrick R, Davies L. Health economics and quality of life in cancer trials: Report based on a UKCCR workshop. Br J Cancer 1998; 77(10): 1543–1548.
3. Kiebert GM, Kaasa S. Quality of life in clinical cancer trials: Experience and perspective of the European organization for research and treatment of cancer. J Nat Cancer Inst Monographs 1996; 20: 91–97.

4. Beitz J, Gnecco C, Justice R. Quality of life end points in cancer clinical trials: The US food and drug administration perspective. J Nat Cancer Inst Monographs 1996; 20: 7–9.
5. EMEA. Recommendations for Efficacy Measurement in Oncology 1995.
6. Moinpour CM, Ganz P, Rowland J, Gritz E, Gotay C. Guest editorial: The value of quality of life data in judging patient benefit: Experts respond to ODAC. The Cancer Letter 1999; 25: 1–5.
7. Tannock IF, Osba D, Stockler MR, et al. Chemotherapy with mitoxantrone plus prednisone or prednisone alone for symptomatic hormone-resistant prostate cancer: A randomized trial with palliative endpoints. J Clin Oncol 1996; 14: 1756–1764.
8. Taylor KM, Macdonald KG, Bezjak A, Ng P, DePetrillo AD. Physicians' perspective on quality of life: An exploratory study of oncologists. Qual Life Res 1995; 5: 5–14.
9. Morris J, Perez D, McNoe B. The use of quality of life data in clinical practice. Qual Life Res 1998; 7: 89–91.

Address for correspondence: G. Kiebert, MEDTAP International, 27 Gilbert Street, London W1Y 1RL, UK
Phone: +44 20 7290 9405; Fax: +44 20 7290 9705
E-mail: kiebert@medtap.co.uk